Journal of Machine Learning Research 66 (2017) 17-33 Submitted 5/17; Published 08/17

Predicting Tomorrow's Mood, Health, and Stress Level using Personalized Multitask Learning and Domain Adaptation

Natasha Jaques JAQUESN@MIT.EDU
Ognjen (Oggi) Rudovic ORUDOVIC@MIT.EDU
Sara Taylor SATAYLOR@MIT.EDU
Akane Sano AKANES@MIT.EDU
Rosalind Picard PICARD@MIT.EDU

Media Lab
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Abstract

Predicting a person's mood tomorrow, from data collected unobtrusively using wearable sensors and smartphones, could have a number of beneficial clinical applications; however, this prediction is an extremely challenging problem. Past approaches often lack the accurate and reliable performance necessary for real-world applications. We posit that this is due to the inability of traditional, one-size-fits-all machine learning models to account for individual differences. To overcome this, we treat predicting tomorrow's mood for a single person as one task, or problem domain. We then adopt Multitask Learning (MTL) and Domain Adaptation (DA) approaches to learn a model which is customized for each person, while still being able to benefit from data across the population. Empirical results on real-world, continuous monitoring data show that the new personalized models — a MTL deep neural network, and a Gaussian Process with DA — both significantly outperform their generic counterparts, providing substantial performance enhancements in automatic prediction of continuous levels of tomorrow's reported mood, stress, and physical health based on data through today.

Keywords: Multitask Learning, Domain Adaptation, Neural networks, Gaussian processes, Affective computing, Mood prediction, Personalization

1. Introduction

Measures of self-reported mood and wellbeing have important ties to clinical health consequences. For example, not only is self-reported unhappiness indicative of scores on clinical measures of depression (Cheng and Furnham, 2003), but self-reported happiness is so strongly associated with longevity that the effect size is comparable to that of cigarette smoking (Veenhoven, 2008). The association between perceived stress and susceptibility to infection and illness has been clearly demonstrated (Cohen et al., 1991). Finally, recent findings from a 29-year study revealed that the single most predictive measure of mortality risk was subjective health status, or self-reported health (Aichele et al., 2016). Clearly, a

machine learning (ML) system that could be trained to predict individuals' self-reported happiness, stress, and health could have a number of important benefits.

In this work we investigate how to build such a ML system, using self-reports of mood and wellbeing collected every day for a month from 70 people. Participants are monitored 24/7 with wrist-worn physiological sensors, a smartphone app, and short daily surveys. This data, as well as historical information about the weather, is used to build personalized deep neural network (DNN) and Gaussian process (GP) models using multitask learning and domain adaptation.

The eventual goal of this work is to construct a system that can unobtrusively monitor participants' data to make predictions about their future mood and wellbeing. Such a system could not only allow users to modify their behavior to promote improved wellbeing, but could also be used to detect early warning signs of depression, anxiety, and mental illness.

While the potential benefits are compelling, constructing a reliable mood prediction system is a challenging task. Despite researchers' best efforts, many previous mood and affect recognition systems — including those trained on the same dataset used for this work — have yielded disappointingly low performance, with accuracies ranging from 55-76% in binary affect classification tasks (e.g. (Bogomolov et al., 2014; Canzian and Musolesi, 2015; Grünerbl et al., 2015; Jaques et al., 2015a,b)). Further, almost all previous work has focused on mood recognition (i.e. detecting the user's current mood), rather than mood prediction. While this is helpful from the perspective of clinical monitoring applications, it does not provide an individual insight into her future mood or give her a chance to make adjustments to improve it.

Another important limitation of existing systems for mood prediction is that mood is treated as a binary state (e.g., happy vs. unhappy). Such a coarse approach can easily miss important distinctions relevant for clinical applications. For example, the lowest possible mood score is treated as equivalent to a slightly lower than average score, while it is possible that only the former is a clinically significant sign of depression. Being able to directly predict a fine-grained estimate of mood, rather than a binary category, could be extremely valuable. Further, since the system relies on continuous collection of data from many sources in real-world, daily life settings, the data will inevitably contain noise. A robust ML system that can provide an estimate of the degree of uncertainty for a given mood prediction is therefore highly desirable.

Perhaps the biggest shortcoming of a typical mood prediction system is that it attempts to predict every person's mood using the same, one-size-fits-all model. Such an approach is inherently limited, due to the high degree of individual variation in how a person's behavior and environment may affect their mood. A number of studies demonstrate that individuals with different personalities may respond differently to the same stimuli; further, personality can strongly affect mood and vulnerability to mental health issues such as depression (Clark et al., 1994). People's moods are even affected differently by the weather (Klimstra et al., 2011). The lack of ability to account for these individual differences may help to explain why many ML models tend to perform poorly in predicting mood. In fact, some authors have already found that personalization can provide important performance enhancements (e.g. (Canzian and Musolesi, 2015; LiKamWa et al., 2013)), although personalization typically takes the form of training many, independent models for each person. However, these

approaches assume an abundance of person-specific data, which in clinical applications is not always the case. Moreover, they fail to leverage the data of all people, which can be used to build more reliable ML models.

We propose principled methods for personalizing ML models, in which we train a customized model for each individual that is still able to benefit from the data of similar people. We begin by training DNNs and GPs to simultaneously predict tomorrow's mood, health and stress intensity from data about their physiology, behavior, and the weather today. Multi-task learning (MTL) is then used to train a personalized DNN, in which person-specific hidden layers are trained to predict the outcome for each individual, but all people are able to benefit from shared feature extraction layers. We use a Domain Adaptation (DA) approach to adapt GPs in a personalized manner, adjusting the models to each person by updating the posterior distribution of the GP. Empirical results demonstrate that the proposed personalization results in considerable performance boost. To the best of our knowledge, this is the first personalized approach for automatic prediction of self-reported mood and wellbeing levels. These personalized, fine-grained predictions of future mood and wellbeing have potential to meaningfully improve real-world monitoring and intervention applications.

2. Related Work

A growing body of work has focused on using data collected unobtrusively via mobile phones and wearable sensors to predict mood and affect. Bogomolov et al. (2014) used mobile phone, weather, and personality data to predict binary stress labels, and achieved 72% classification accuracy. Grünerbl et al. (2015) used long-term smartphone monitoring to detect depressive and manic states in bipolar disorder, obtaining 76% prediction accuracy. Previous work using the same dataset discussed in this paper explored a variety of methods for predicting mood, obtaining prediction accuracies ranging from 62-74% (Jaques et al., 2015a,b). Nonetheless, none of these works addressed prediction of mood and wellbeing levels, but only their binary classification.

Other approaches to mood and health estimation have looked specifically at the effect of personalization, in the context of training separate, individual models for each person. Canzian and Musolesi (2015) found that when using location data to detect depressed mood, training a generic SVM over all people yielded worse sensitivity and specificity scores (0.74 and 0.78, respectively), than training a separate SVM for each person (0.71 and 0.87). Similarly, Clifton et al. (2013) found that training an individual GP per person allowed for better prediction from wearable sensor data.

In LiKamWa et al. (2013), researchers used a smartphone app to infer user mood based on communication history and usage patterns, and found that a generic model could only classify mood with 66% accuracy. However, if two months of longitudinal data were obtained for each person and used to train a person-specific individual classifier, accuracy could improve to 93%! The authors also explored a simple hybrid approach for combining generic and personal classifiers when there is not enough data to train person-specific models. Since our data does not contain enough days per person (~ 25) to train individual DNNs or GPs, we extend this work by proposing principled ways for building personalized models that can leverage the data of all individuals.

3. Methods

Traditional ML approaches attempt to learn a single function over all of the available data. We hypothesize that such approaches are inappropriate for affective computing tasks such as mood prediction, due to their inability to account for individual differences. People can show profound differences in their affective response to the same stimuli; for example, an introvert attending a loud, crowded party may exhibit a very different stress response than an extrovert (Brebner, 1990).

While this challenge could potentially be solved by training separate models for each person, in many affective computing tasks gathering data is extremely expensive. Also, datasets are typically quite small, rendering training powerful models with this approach infeasible (as is the case with our data). Further, a more robust model can be trained by intelligently learning from the data of other people to the degree that it is relevant to each individual. Therefore, we propose a personalized modeling approach in which data of each person are treated as a separate mood prediction domain or task. To benefit from available data of all persons during the learning stage, we customize the proposed models for each person using Domain Adaptation (DA) or Multitask Learning (MTL) techniques, as described below.

3.1 Gaussian Processes for Personalized DA

We consider a supervised setting for domain adaptation, where we are given a relatively large amount of labeled training data (*source* domain), and a considerably smaller set of labeled data in the *target* domain. Furthermore, we assume a person-dependent setting, i.e., the (non-overlapping) data of target persons are available in the source and target domain. Thus, our goal is to learn a general prediction model from data of all persons, and then leverage the limited data of a target person to perform the model adaptation to that specific person. Formally, let $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ be the input (features) and output (labels) spaces, respectively. We assume that the input space is composed of the source and target domains, $\mathcal{S}$ and $\mathcal{T}$, respectively, that may differ in feature distribution. Hence, $\boldsymbol{X}^{(s)} = \{\boldsymbol{x}_{n_s}^{(s)}\}_{n_s=1}^{N_s}$ and $\boldsymbol{X}^{(t)} = \{\boldsymbol{x}_{n_t}^{(t)}\}_{nt=1}^{N_t}$, with $\boldsymbol{x}_{n_s}^{(s)}, \boldsymbol{x}_{n_t}^{(t)} \in \mathbb{R}^D$, and $N_t \ll N_s$. Similarly, $\boldsymbol{Y}^{(s)} = \{\boldsymbol{y}_{n_s}^{(s)}\}_{n_s=1}^{N_s}$ and $\boldsymbol{Y}^{(t)} = \{\boldsymbol{y}_{n_t}^{(t)}\}_{nt=1}^{N_t}$ are the output labels for the source and target domains, respectively, where $\boldsymbol{y}_n^{\{s,t\}}$ represents the intensity level of the wellbeing dimension that we wish to estimate (i.e., mood, health, or stress).

3.1.1 Gaussian Processes (GPs)

Here, we introduce briefly the modeling framework of GP regression, that we use as the base model in our personalization approach via domain adaptation. We employ GPs for two reasons: (i) it is a non-parametric model, which allows us to efficiently capture non-linear relationships between input features and output labels using kernel functions; and (ii) due to its probabilistic nature, the model adaptation can be performed in a principled manner by deriving a posterior distributions conditioned on the adaptation (target) data. The GP regression function is defined as:

$$\boldsymbol{y}_{n_v}^{(v)} = f^{(v)}(\boldsymbol{x}_{n_v}^{(v)}) + \epsilon^{(v)}, \tag{1}$$

where $\epsilon^{(v)} \sim \mathcal{N}(0, \sigma_v^2)$ is i.i.d. additive Gaussian noise, and the index $v \in \{s, t\}$ denotes the dependence on each domain. While in a traditional GP, all data is considered to come from the same domain, in our DA approach we focus on adapting models to new domains. The objective of a GP is to infer the latent functions $f^{(v)}$, given the training dataset $\mathcal{D}^{(v)} = \{\boldsymbol{X}^{(v)}, \boldsymbol{Y}^{(v)}\}$. By following the framework of GPs (Rasmussen and Williams, 2006), we place a prior on the functions $f^{(v)}$, so that the function values $\boldsymbol{f}_{n_v}^{(v)} = f^{(v)}(\boldsymbol{x}_{n_v}^{(v)})$ follow a Gaussian distribution $p(\boldsymbol{F}^{(v)}|\boldsymbol{X}^{(v)}) = \mathcal{N}(\boldsymbol{F}^{(v)}|\boldsymbol{0}, \boldsymbol{K}^{(v)})$. Here, $\boldsymbol{F}^{(v)} = \{\boldsymbol{f}_{n_v}^{(v)}\}_{n_v=1}^{N_v}$, and $\boldsymbol{K}^{(v)} = k^{(v)}(\boldsymbol{X}^{(v)}, \boldsymbol{X}^{(v)})$ is the kernel covariance function. We use the radial basis function (RBF) kernel, defined as:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\Big(-\frac{1}{2\ell^2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2\Big), \tag{2}$$

where $\{\ell, \sigma_f\}$ are the kernel hyper-parameters. The regression function is then fully defined by the set of hyper-parameters (hp) $\boldsymbol{\theta} = \{\ell, \sigma_f, \sigma_v\}$. Training of the GP consists of finding the hp that maximize the log-marginal likelihood:

$$\begin{aligned}\log p(\boldsymbol{Y}^{(v)}|\boldsymbol{X}^{(v)}, \boldsymbol{\theta}^{(v)}) = &- \operatorname{tr}\Big[(\boldsymbol{K}^{(v)} + \sigma_v^2\boldsymbol{I})^{-1}\boldsymbol{Y}^{(v)}\boldsymbol{Y}^{(v)^T}\Big] \\ &- \log|\boldsymbol{K}^{(v)} + \sigma_v^2\boldsymbol{I}| + \text{const.}\end{aligned} \tag{3}$$

Given a test input $\boldsymbol{x}_*^{(v)}$ we obtain the GP predictive distribution by conditioning on the training data $\mathcal{D}^{(v)}$ as $p(\boldsymbol{f}_*^{(v)}|\boldsymbol{x}_*^{(v)}, \mathcal{D}^{(v)}) = \mathcal{N}(\mu^{(v)}(\boldsymbol{x}_*^{(v)}), V^{(v)}(\boldsymbol{x}_*^{(v)}))$ with

$$\mu^{(v)}(\boldsymbol{x}_*^{(v)}) = \boldsymbol{k}_*^{(v)^T}(\boldsymbol{K}^{(v)} + \sigma_v^2\boldsymbol{I})^{-1}\boldsymbol{Y}^{(v)} \tag{4}$$

$$V^{(v)}(\boldsymbol{x}_*^{(v)}) = k_{**}^{(v)} - \boldsymbol{k}_*^{(v)^T}(\boldsymbol{K}^{(v)} + \sigma_v^2\boldsymbol{I})^{-1}\boldsymbol{k}_*^{(v)}, \tag{5}$$

where $\boldsymbol{k}_*^{(v)} = k^{(v)}(\boldsymbol{X}^{(v)}, \boldsymbol{x}_*^{(v)})$ and $k_{**}^{(v)} = k^{(v)}(\boldsymbol{x}_*^{(v)}, \boldsymbol{x}_*^{(v)})$. For convenience we denote $\boldsymbol{\mu}_*^{(v)} = \mu^{(v)}(\boldsymbol{x}_*^{(v)})$ and $V_{**}^{(v)} = V^{(v)}(\boldsymbol{x}_*^{(v)})$. In most applications, the GP mean function is used as the point estimate of the output targets. However, this is a generic model, i.e., it is not optimized to achieve the best performance on each target person. We describe below the adaptation approach based on GPs that we devise for personalized estimation of mood, health and stress levels for each target person.

3.1.2 GP Adaptation with Posterior Adaptation

The probabilistic nature of GPs allows us to easily incorporate new data of target person into the model, without re-training the model. This results in a posterior distribution of GPs, rendering a personalized regression function specifically tuned to the target person. To this end, we exploit the Bayesian adaptation approach proposed in (Liu and Vasconcelos, 2015; Eleftheriadis et al., 2016). The GP model adaptation consists of the following three key components:

1. A GP trained on the source data with likelihood $p(\boldsymbol{Y}^{(s)}|\boldsymbol{X}^{(s)}, \boldsymbol{\theta})$ and hp $\boldsymbol{\theta}$ is trained as a base model, and is defined by Eqs. (4–5).

2. The posterior distribution of the base GP model is then used as a prior for the GP evaluated on target (adaptation) data $p(\boldsymbol{Y}^{(t)}|\boldsymbol{X}^{(t)}, \mathcal{D}^{(s)}, \boldsymbol{\theta})$.

3. The posterior distribution over the target data is then corrected to account for the adaptation data $\mathcal{D}^{(t)}$ of the target person.

The prior over the target data in the second step is given by applying Eqs. (4–5) on $\boldsymbol{X}^{(t)}$ as:

$$\boldsymbol{\mu}^{(t|s)} = \boldsymbol{K}_{st}^{(s)^T}(\boldsymbol{K}^{(s)} + \sigma_s^2\boldsymbol{I})^{-1}\boldsymbol{Y}^{(s)} \tag{6}$$

$$\boldsymbol{V}^{(t|s)} = \boldsymbol{K}_{tt}^{(s)} - \boldsymbol{K}_{st}^{(s)^T}(\boldsymbol{K}^{(s)} + \sigma_s^2\boldsymbol{I})^{-1}\boldsymbol{K}_{st}^{(s)}, \tag{7}$$

where $\boldsymbol{K}_{tt}^{(s)} = k^{(s)}(\boldsymbol{X}^{(t)}, \boldsymbol{X}^{(t)}), \boldsymbol{K}_{st}^{(s)} = k^{(s)}(\boldsymbol{X}^{(s)}, \boldsymbol{X}^{(t)})$, and the superscript $t|s$ denotes the conditioning order. Given the above prior and a test input $\boldsymbol{x}_*^{(t)}$ (i.e., features of target person), the correct form of the adapted posterior after observing the target person adaptation data is given by:

$$\mu_{ad}^{(s)}(\boldsymbol{x}_*^{(t)}) = \boldsymbol{\mu}_*^{(s)} + \boldsymbol{V}_*^{(t|s)^T}(\boldsymbol{V}^{(t|s)} + \sigma_s^2\boldsymbol{I})^{-1}(\boldsymbol{Y}^{(t)} - \boldsymbol{\mu}^{(t|s)}) \tag{8}$$

$$V_{ad}^{(s)}(\boldsymbol{x}_*^{(t)}) = V_{**}^{(s)} - \boldsymbol{V}_*^{(t|s)^T}(\boldsymbol{V}^{(t|s)} + \sigma_s^2\boldsymbol{I})^{-1}\boldsymbol{V}_*^{(t|s)}, \tag{9}$$

with $\boldsymbol{V}_*^{(t|s)} = k^{(s)}(\boldsymbol{X}^{(t)}, \boldsymbol{x}_*^{(t)}) - k^{(s)}(\boldsymbol{X}^{(s)}, \boldsymbol{X}^{(t)})^T(\boldsymbol{K}^{(s)} + \sigma_s^2\boldsymbol{I})^{-1}k^{(s)}(\boldsymbol{X}^{(s)}, \boldsymbol{x}_*^{(t)})$. Our personalized GP model based on posterior adaptation to the target person data (DA-GP) is fully defined by Eqs. (8–9). Note that that final personalized prediction is a combination of the generic (base) model based on the source data only, and a correction term, which shifts the GP mean toward the feature distribution of the target person, while reducing the model's uncertainty in the estimated output. In this way, the model automatically adapts to the range of, for instance, stress levels specific to the target person.

3.2 Deep Neural Networks for Personalized MTL

Multitask learning is a type of transfer learning in which models are learned simultaneously for a set of related tasks, and was originally proposed as a way to induce efficient internal representations in neural networks (NNs) (Caruana, 1997). MTL can also be seen as a powerful form of regularization that can enhance generalization performance, as long as the tasks are sufficiently related (Rosenstein et al., 2005). In the context of NNs, MTL forces the model to learn an efficient and robust internal representation that generalizes well to all of the related tasks.

Our approach is to build a personalized MTL-NN in which predicting the mood of each person is treated as one task. The MTL-NN contains several initial hidden layers that are shared among all the tasks, followed by smaller, final hidden layers that are unique to each task; Figure 1 shows a simplified version of this architecture. The intuition behind this design is that the shared layers will learn to extract information that is useful for summarizing relevant characteristics of any person's day into an efficient, generalizable embedding. The final, task-specific layers are then expected to learn how to map this embedding to a prediction customized for each person based on their typical reactions to a given type of day. For example, if the first layers learn to condense all of the relevant smartphone app data about phone calls and texting into an aggregate measure of social support, the task-specific layers can then learn a unique weighting of this measure for each

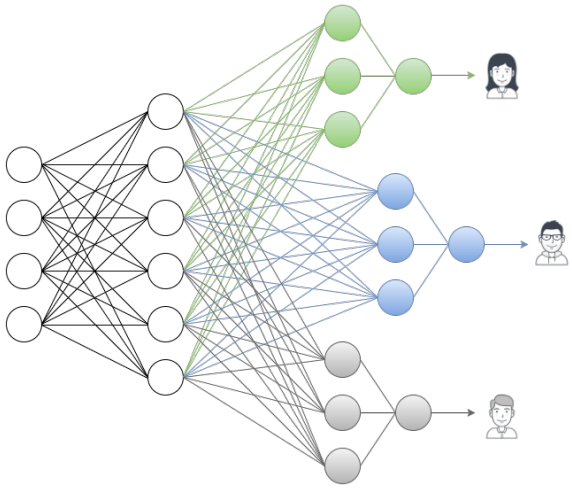


Figure 1: A simplified version of the MTL-NN architecture, in which each person has a unique, task-specific layer.

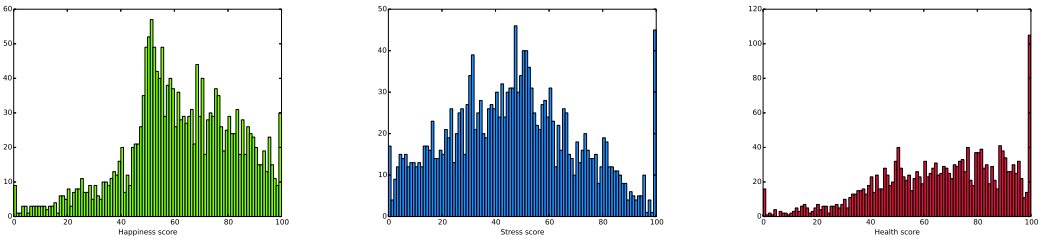
person; perhaps Person A is more strongly affected by a lack of social support than Person B.

To train the network, each batch of training data must be selected from a single person. This batch is then used to predict the target labels for that person, and the errors are backpropagated to update that person's task-specific weights, as well as adjust the weights within the shared layers. By continuing to randomly sample people and update both the person-specific and shared weights, the network will eventually learn a shared representation relevant to each and every person.

While deep learning is a powerful branch of ML, when training on small datasets it is important to heavily regularize the network to avoid overfitting. Although MTL itself is a strong form of regularization, we also impose a penalty on the L2 norm of the network's weights, and train the network to simultaneously predict all 3 wellbeing labels (mood, stress, and health) to further improve the generalizability of the embedding. We also apply dropout, a popular approach to NN regularization that is equivalent to training an ensemble of NNs which share parameters on bagged samples of the data (Srivastava et al., 2014).

4. Mood prediction dataset

We employ a dataset collected as part of a long-term, collaborative study between MIT and Brigham and Women's hospital (Sano, 2015). The study, entitled SNAPSHOT (Sleep, Networks, Affect, Performance, Stress, and Health using Objective Techniques) collected extremely rich data from 206 undergraduate students, who were monitored 24/7 for 30 days each using a combination of physiological sensors, a smartphone app, and behavioral surveys. Along with weather data obtained from DarkSky's Forecast.io API (LLC, 2016), the SNAPSHOT data were used to extract a total of 343 features at the granularity of one day, which pertain to each participants' physiology, phone usage, behaviors, location, and of course, the weather. We briefly describe these features.



(a) Mood (b) Stress (c) Health

Figure 2: Distribution of self-report labels in the data. Students frequently report feeling healthy and stressed.

4.1 Physiology

Physiological data, including skin conductance, skin temperature, and accelerometer data were recorded throughout the day at an 8Hz sampling rate using Affectiva Q sensors. Participants' steps and stillness (periods during which they were inactive) were computed, since there are obvious associations between exercise, sedentary activity, and mental health and wellbeing (Salmon, 2001). Features related to skin temperature over different time intervals of the day were also extracted, since they may be relevant to the body's circadian rhythm and thus to wellbeing (Partonen, 1996).

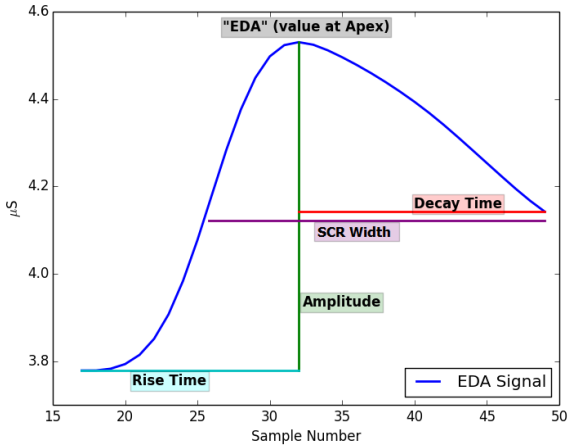


Figure 3: Features extracted for each detected SCR.

Skin conductance (SC) is used to measure electrical changes across the surface of the skin, which are a reflection of sympathetic nervous system activity without parasympathetic antagonism; the skin conductance can increase in response to temperature, exercise, uncertainty and anticipation, strong emotions, and some kinds of stress (Boucsein, 2012). To extract useful information from this signal, we first detect skin conductance responses (SCRs), peaks in the signal that indicate increased SC activity. We then use a previously trained ML classifier (Taylor et al., 2015) to detect and remove any SCRs that may actually

be recording artifacts, which occur due to the noisy, ambulatory nature of the data collection. The remaining SCRs were then used to compute a number of features related to their magnitude and shape, as shown in Figure 3. We also compute SC signals weighted by the inverse of skin temperature and motion, which are stronger when SC is high in the absence of exercise or temperature increase; therefore, this signal is more likely to relate to cogitive or emotional arousal. In total, we compute 172 features related to SC, skin temperature, and motion.

4.2 Location

Students' GPS coordinates are logged throughout the day via the smartphone app. After cleaning and filtering the raw data, 15 features such as the maximum distance traveled per day, time spent on the university campus, and time spent outdoors are computed. To more effectively model students' location patterns, we learn a Gaussian Mixture Model (GMM) over the history of their coordinates (see Figure 4), and use this model to compute features such as the log likelihood of a given day. This feature will be higher if the day is more routine, a factor that previous work has shown to be negatively associated with happiness and calmness (Jaques et al., 2015a).

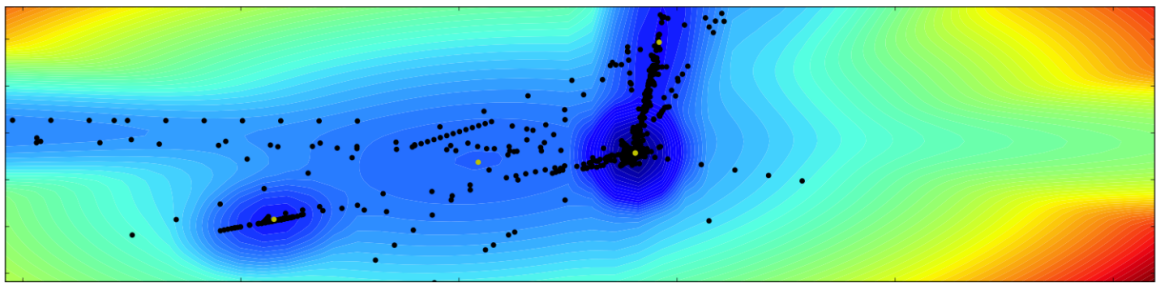


Figure 4: Probability distribution learned by a GMM over one participant's location patterns. Black dots are logged coordinates, blue contours describe the most likely locations

4.3 Phone

An app on participants' phones logs their text messages (SMS), calls, and screen on/off events throughout the day. These data are used to compute 20 call, 30 SMS, and 25 screen features related to the duration and timing of these events. For both SMS and call, the features include the number of unique contacts with whom each participant interacts per day, since social support is strongly related to wellbeing (Cohen and Wills, 1985). Both the phone and physiology features are computed over four time intervals per day: 12-3AM, 3-10AM, 10AM-5PM, 5-11:59PM. These intervals were arrived at by examining density plots of the times students were most likely to be asleep (3-10AM), or in class (10AM-5PM).

4.4 Surveys

At the same time that students report their mood and wellbeing, they report on their sleep, exercise, academic and extra-curricular activities, social interactions, and alcohol and drug

consumption. In addition to these features, we include the day of the week and whether there is school the next day (whether it is a school night), for a total of 38 features.

4.5 Weather

The effect of weather on mood is has been validated in a number of studies (e.g. (Partonen, 1996; Li et al., 2014)). We compute 40 features based on weather data that relate to sunlight, temperature, wind, Barometric pressure, and the difference between today's weather and the rolling average.

4.6 Mood and Wellbeing Labels

Participants reported their wellbeing daily, in terms of their mood (sad/happy), stress (calm, relaxed/stressed), and health (sick/healthy) on an sliding scale, resulting in ordinal wellbeing scores ranging from 0-100. Figure 2 shows the distribution of self-report labels in the data. Previous work on this dataset has approached detecting or predicting mood as a classification problem, splitting these continuous scores into binary labels (Jaques et al., 2015a,b, 2016). Since mood classification is such a difficult problem, authors of previous work chose to discard the middle 40% of scores, comprising the most neutral wellbeing scores (i.e. those most near the median), in order to disambiguate the classification labels. In contrast, we use regression techniques to directly predict the ordinal label values that were reported.

5. Experiments

To ensure that we are able to train robust personalized prediction models, we restrict our attention to only those participants who provided at least 25 days of data in which all data sources are present. Since the data is noisy, this reduces the dataset to a total of 69 participants and 1895 days worth of data. These samples are then divided into non-overlapping training, validation, and testing sets using a 60/20/20% split.

The personalized models were compared to their single-task learning (STL) counterparts, a standard GP and NN. The code for the GP and NN models was written using gpml Matlab code[1] and Tensorflow, respectively. The GP model hyper-parameters were learned using conjugate-gradient optimization (Rasmussen and Williams, 2006). For the adaptation setting, we used the validation data of each target person.

In training both the STL and MTL NN, a grid search was used to select hyperparameters (including the number and size of hidden layers) by assessing performance on the validation set. By initializing the MTL NN with the pre-trained weights of the STL NN, we found we could successfully train a much higher capacity personalized network. In the end, we found that an architecture of four hidden layers with sizes 2048, 1024, 512, and 256, a dropout factor of 0.25, and no L2 regularization gave the best performance for the STL NN, and we therefore adopted the same architecture for the MTL-NN. However, we found that while a batch size of 32 was effective for training the STL NN, a batch size of 1 gave the best results for training the MTL-NN. This could be because minibatch updates are only an unbiased

1. http://www.gaussianprocess.org/gpml/code/

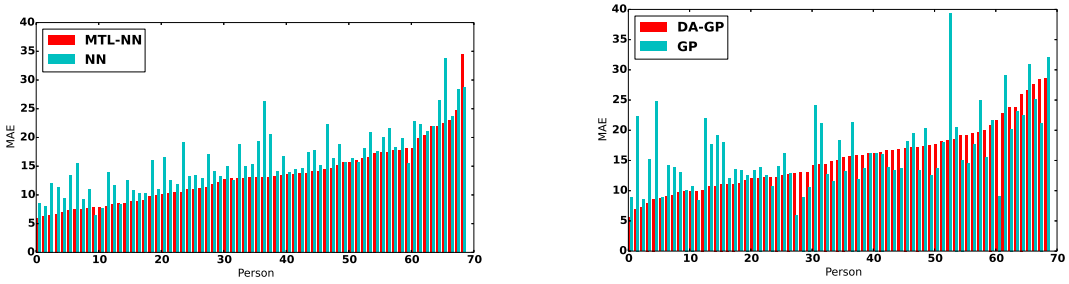
estimate of the true gradient when the samples within a minibatch are uncorrelated, which necessitates a batch size of 1 in the MTL setting.

6. Results

The Mean Absolute Error (MAE) obtained for both the traditional and personalized models is shown in Table 1. While both the MTL-NN and DA-GP offer reduced MAE in predicting each of the 3 wellbeing scores, paired-samples t-tests revealed that these differences are only consistently significantly different for the MTL-NN model. This difference could be due to the fact that the MTL-NN model has implicit knowledge of each of the participants' ID, due to the way the model is constructed. In contrast, the DA-GP must learn which samples are most similar to each person, which constitutes a more difficult problem. Further, the small number of samples per person in the validation set (~ 5 in this case) makes it difficult for the DA-GP to perform a robust posterior correction (Eqs.(6-7)). Note that we also attempted learning of the base GP model using both training and validation data, however, the results only marginally improved over the GP trained using training data only. This clearly shows the benefits of the proposed personalized GP adaptation scheme.

	Model	Mood	Stress	Health	Total
Traditional	GP	16.0	17.2	16.7	16.6
	NN	15.0	17.1	16.5	16.2
Personalized	DA-GP	14.8	16.4	**14.6**	15.3
	MTL-NN	**13.0**	**14.1**	**12.9**	**13.3**

Table 1: MAE in predicting wellbeing on the held-out test set, for both traditional ML models and personalized models. Bolded entries represent significant improvements over the non-personalized version of the model ($p < .05$).



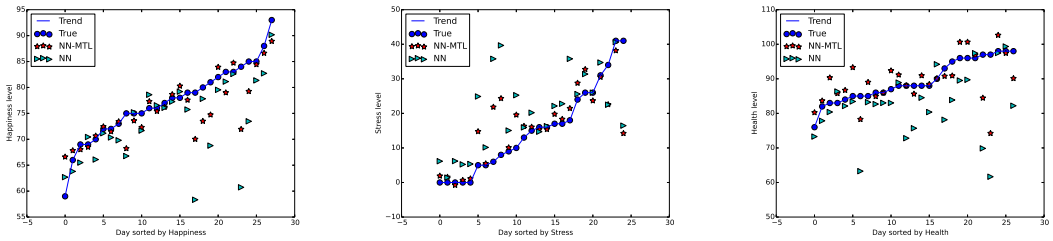
(a) NN models (b) GP models

Figure 5: MAE per person for both the personalized and generic models. For 61 out of 69 participants, MAE is lower with the personalized MTL-NN model. For GP's personalization is better for 40 out of 69 participants.

Figure 5 shows the MAE for each person. For 61/69 people, the personalized MTL-NN provides lower error than the generic NN. While this effect is not as strong for the DA-GP, having a personalized model still benefits the majority of people. Clearly, personalization

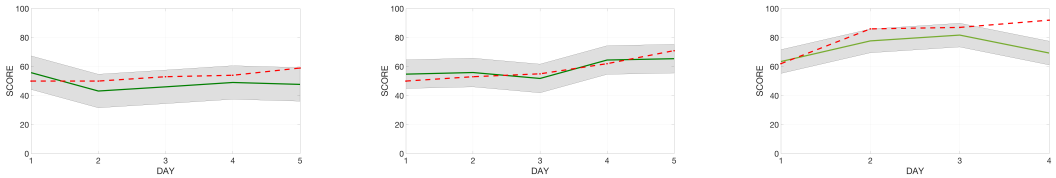
can not only provide performance advantages across all participants, but it ensures that there are fewer people for whom the model cannot make accurate predictions.

Figure 6 shows the actual predictions of the NN and MTL-NN on each outcome label for three randomly selected participants. As is evident in the figure, the MTL-NN is able to provide a close fit to the ground truth data. Figure 6 also helps to demonstrate the degree of individual variability within the data; while the average Health report was 65.60 (SD=23.08), the participant in Figure 6 (c) only reports Health scores ranging from 76-98. This could explain why the predictions of the non-personalizd NN are so frequently drawn downwards toward the group average; it has no ability to learn to adapt its predictions to this participants' unique health pattern. Similarly, Figure 7 shows the predicted mean and variance of the DA-GP over the held-out test data.



(a) Mood (b) Stress (c) Health

Figure 6: NN and MTL-NN predictions for each outcome for a randomly selected subject compared to the ground truth mood report data, which has been sorted by intensity and connected with a trend line.



(a) Mood (b) Stress (c) Health

Figure 7: Predicted mean and variance over the held-out test data as learned by the DA-GP. The lines in red and green represent the reported and predicted scores, respectively. As can be seen from depicted certainty levels (in gray), the predicted values are within one standard deviation (uncertainty) intervals estimated by GP.

In addition to assessing the absolute differences between the model's predictions and the ground truth, we are interested in determining if the models are able to capture the true trends underlying participants' mood and wellbeing levels; in other words, which models provide an overall better fit to the data? We assess this using the Intraclass Correlation Coefficient (ICC(3,1)), which not only measures the association between the model's predictions and the true ratings (as in Pearson's correlation), but also penalizes absolute distance from the ground truth. We find that the performance benefits offered by the personalized MTL-NN and DA-GP are substantially more pronounced in terms of ICC, as shown in

Table 2. In this case, the benefit of the DA-GP is clearly apparent, as it provides a 100% improvement above the generic GP baseline. This is likely because the adjustment made to adapt the mean and variance of the posterior distribution for each person allows the DA-GP to adapt its predictions toward each person's true mean mood value. This dramatic improvement in ICC therefore suggests that the personalized models can more accurately distinguish the participants that have the lowest mood and health, and the highest stress - an important ability for most clinical applications.

	Model	**Mood**	**Stress**	**Health**	**Total**
	GP	.176	.358	.286	.274
Traditional	NN	.262	.422	.373	.352
	DA-GP	**.461**	**.587**	**.606**	**.551**
Personalized	MTL-NN	**.441**	**.621**	**.613**	**.558**

Table 2: ICC, a measure of model fit, in predicting wellbeing on the held-out test set. Bolded entries indicate an improvement of at least 50% over the non-personalized model.

6.1 Limitations and Future Work

Clearly, personalization can provide advantages in predicting mood and wellbeing, a problem where interindividual variability is high. However, it is important to note that by their nature, these models are person-dependent; that is, they require at least some labeled data from each person in order to be trained. While the DA and MTL approaches discussed here provide the advantage that they can be trained even when there is not enough data per-person to train many individual person-specific models, in our case we still require at least 15 days of training data per person. This implies that a new user of a mood prediction system built using these models would have to input their mood and wellbeing for 15 days to obtain the level of performance presented here.

While asking users to report their mood for roughly two weeks is not unreasonable, given that most users of a quantified-self device or system continue to use it for about six months (Patel et al., 2015), it is important to note that simple extensions to the models can allow them to make predictions in the absence of any self-reports from a given user. For example, the MTL-NN could make predictions for a new user by feeding the data through each output head and averaging the predictions; we would expect this approach to recover approximately the performance of the impersonal NN model. For GPs, adapting to a new person is even simpler. In the case of the DA-GP posterior model, with only a few days of labeled data (e.g. 5, in this case), it is possible to notably improve the predictions. It is not necessary that a person's data be part of the training set to perform the adaption. Also, multi-task extensions of GPs are another venue to pursue when personalizing models, in order to take advantages of the MTL paradigm.

7. Conclusions

This work has empirically demonstrated that the performance of machine learning mood prediction systems can be meaningfully enhanced by personalizing those models in a principled way. We have outlined two methods for accomplishing personalization; by using

a Domain Adaptation approach to adapt the posterior distribution of the model towards each person's unique mood and wellbeing level, and by using Multitask Learning to train a deep neural network with specialized final layers for each person. Not only do these models provide 13-22% lower average error than traditional models in making fine-grained predictions about participants' outcomes, but we find that the personalized models provide a significantly better fit to the data, improving ICC by as much as 160% above the generic GP baseline. These performance improvements may have important clinical benefits, such as enabling a model to better distinguish between participants who are severely unhappy or stressed, which then can enable more relevant and targeted treatments for improving wellbeing.

Acknowledgments

We would like to thank Dr. Charles Czeisler, Dr. Elizabeth Klerman, Conor O'Brien and other SNAPSHOT project members for their help in running the SNAPSHOT study. This work was supported by the MIT Media Lab Consortium, NIH Grant R01GM105018, Samsung Electronics, and Canada's NSERC program. The work of O. Rudovic is funded by European Union H2020, Marie Curie Action - Individual Fellowship no. 701236 (EngageMe).

References

S. Aichele, P. Rabbitt, and P. Ghisletta. Think fast, feel fine, live long: A 29-year study of cognition, health, and survival in middle-aged and older adults. *Psychological science*, 27 (4):518–529, 2016.

A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. Pentland. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *ICME*, pages 477–486. ACM, 2014.

W. Boucsein. *Electrodermal activity*. Springer Science+Business Media, LLC, 2012.

John Brebner. Personality factors in stress and anxiety. *Cross-cultural anxiety*, 4:11–19, 1990.

L. Canzian and M. Musolesi. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Pervasive and Ubiquitous Computing*, pages 1293–1304. ACM, 2015.

R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

H. Cheng and A. Furnham. Personality, self-esteem, and demographic predictions of happiness and depression. *Personality and individual differences*, 34(6):921–942, 2003.

L. Clark, D. Watson, and S. Mineka. Temperament, personality, and the mood and anxiety disorders. *Journal of abnormal psychology*, 103(1):103, 1994.

L. Clifton, D. Clifton, M. Pimentel, P. Watkinson, and L. Tarassenko. Gaussian processes for personalized e-health monitoring with wearable sensors. *IEEE Transactions on Biomedical Engineering*, 60(1):193–197, 2013.

S. Cohen and T. Wills. Stress, social support, and the buffering hypothesis. *Psychological bulletin*, 98(2):310, 1985.

S. Cohen, D. Tyrrell, and A. Smith. Psychological stress and susceptibility to the common cold. *New England journal of medicine*, 325(9):606–612, 1991.

S. Eleftheriadis, O. Rudovic, M. Deisenroth, and M. Pantic. Gaussian process domain experts for model adaptation in facial behavior analysis. In *CVPR'W*, pages 18–26, 2016.

A. Grünerbl, A. Muaremi, V. Osmani, G. Bahle, S. Oehler, G. Tröster, O. Mayora, C. Haring, and P. Lukowicz. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE J. of Biomed. H. Info.*, 19(1):140–148, 2015.

N. Jaques, S. Taylor, A. Azaria., A. Ghandeharioun, A. Sano, and R. Picard. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In *ACII*. IEEE, 2015a.

N. Jaques, S. Taylor, A. Sano, and R. Picard. Multi-task, multi-kernel learning for estimating individual wellbeing. In *NIPS 2015 Workshop on Multimodal Machine Learning*, volume 898, 2015b.

N. Jaques, S. Taylor, E. Nosakhare, A. Sano, and R. Picard. Multi-task learning for predicting health, stress, and happiness. In *Proc. NIPS'W on ML for Healthcare.*, 2016.

T. Klimstra, T. Frijns, L. Keijsers, J. Denissen, Q. Raaijmakers, M. van Aken, H. Koot, P. van Lier, and W. Meeus. Come rain or come shine: individual differences in how weather affects mood. *Emotion*, 11(6):1495, 2011.

J. Li, X. Wang, and E. Hovy. What a nasty day: Exploring mood-weather relationship from twitter. In *Int'l Conf. on Info. and Knowledge Management*, pages 1309–1318. ACM, 2014.

R. LiKamWa, Y. Liu, N. Lane, and L. Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *MobiSys*, pages 389–402. ACM, 2013.

B. Liu and N. Vasconcelos. Bayesian model adaptation for crowd counts. In *ICCV*, pages 4175–4183, 2015.

The Dark Sky Company LLC. Dark sky forecast api, 2016. URL `https://developer.forecast.io/`.

T. Partonen. Dopamine and circadian rhythms in seasonal affective disorder. *Med. hypotheses*, 47(3):191–192, 1996.

M. Patel, D. Asch, and K. Volpp. Wearable devices as facilitators, not drivers, of health behavior change. *Jama*, 313(5):459–460, 2015.

C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, MA, 2006.

M. Rosenstein, Z. Marx, L. Kaelbling, and T. Dietterich. To transfer or not to transfer. In *NIPS 2005 Workshop on Transfer Learning*, volume 898, 2005.

P. Salmon. Effects of physical exercise on anxiety, depression, and sensitivity to stress: a unifying theory. *Clinical psychology review*, 21(1):33–61, 2001.

A. Sano. *Measuring College Students Sleep, Stress and Mental Health with Wearable Sensors and Mobile Phones*. PhD thesis, MIT, 2015.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014.

S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, and R. Picard. Automatic identification of artifacts in electrodermal activity data. In *EMBC*. IEEE, 2015.

R. Veenhoven. Healthy happiness: Effects of happiness on physical health and the consequences for preventive health care. *Journal of happiness studies*, 9(3):449–469, 2008.