Supplementary Material of
Active Learning for Decision-Making from Imbalanced Observational Data

Iiris Sundin [1] **Peter Schulam** [*2] **Eero Siivola** [*1] **Aki Vehtari** [1] **Suchi Saria** [2] **Samuel Kaski** [1]

1. Effect of Imbalance on Type S Error Rate

In this section we prove that, under certain assumptions, imbalance increases the error rate in decision-making. We start by a sketch of the proof and then continue with details.

Sketch of proof: First, we assume a probabilistic model of potential outcomes, with broad prior distributions. This implies that when the sample size is small, posteriors will be wide. Then, we show that imbalance decreases the expected number of samples locally, therefore increasing the expected Type S error rate locally. Finally, we provide conditions under which local increase in the expected Type S error rate also increases the expected global Type S error rate.

Assumption S1. (Prior). Assume a broad prior on the expected potential outcomes μ_a: $p(\mu_a) > D > 0 \; \forall \mu_a \in [-K, K]$. The action $a \in \{0, 1\}$.

Assumption S2. (Likelihood). Likelihood of observation $p(y_a \mid \mu_a) > C > 0 \; \forall y_a \in [-K, K]$.

Comment. Consequence of Assumptions S1 and S2 is that if sample size is small, the posterior will be wide.

Lemma S1. Given observations on two potential outcomes $\mathcal{D} = \{y_{1,i}\}_{i=1}^{n_1} \cup \{y_{0,j}\}_{j=1}^{n_0}$, the probability of Type S error has lower bound $p(\text{"Type S error"}) > 2K^2 D^2 C^{n_1+n_0}$.

Proof: We prove Lemma S1 for the case where the true treatment effect is negative, that is, $m_0 > m_1$.

Posterior of μ_a given data is:
$p(\mu_a \mid \mathcal{D}) = \frac{1}{Z} p(y_a \mid \mu_a) p(\mu_a)$.

The probability of the Type S error is

$$
\begin{aligned}
p(\text{"Type S error"}) &= \int_{\mu_0 \leq \mu_1 | m_0 > m_1} p(\mu_1, \mu_0 \mid \mathcal{D}) d\mu_0 d\mu_1 \\
&= \int_{\mu_1} \int_{\mu_0 \leq \mu_1} p(\mu_1 \mid \mathcal{D}) p(\mu_0 \mid \mathcal{D}) d\mu_0 d\mu_1 \qquad (1) \\
&\geq \int_{\mu_1} \int_{\mu_0 \leq \mu_1} \prod_{a=0}^{1} \prod_{i=1}^{n_a} p(y_{a,i} \mid \mu_a) p(\mu_a) d\mu_0 d\mu_1 \\
&\geq \int_{-K}^{K} \int_{-K}^{\mu_1} C^{n_1} D C^{n_0} D d\mu_0 d\mu_1 \qquad (2) \\
&= D^2 C^{n_1+n_0} \int_{-K}^{K} (\mu_1 + K) d\mu_1 \\
&= D^2 C^{n_1+n_0} 2K^2,
\end{aligned}
$$

where (1) follows from assuming factorization, and (2) by Assumptions S1 and S2. □

Assumption S3. (Covariate distributions). Let $p^a(x) := p(x \mid a)$ be the covariate distribution of group a in covariate space $\mathcal{X}$. Assume $p^a(x)$ are Lipschitz continuous with constant L.

Definition: (Imbalance). Imbalance can be measured using Integral Probability Metric as described by Shalit et al. (2017). Let G be a function family consisting of functions $g : \mathcal{X} \to \mathbb{R}$. For a pair of distributions p, q over $\mathcal{X}$ the Integral Probability Metric is defined as

$$
IPM_G(p, q) = \sup_{g \in G} \left| \int_{\mathcal{X}} g(x)(p(x) - q(x)) dx \right|. \qquad (3)
$$

Assumption S4. (Imbalance). Assume there exists non-empty $\Omega = \{x \in \mathcal{X} \mid |p^1(x) - p^0(x)| \geq h\}$ where $h > 0$. (For small enough h this holds if there is any imbalance).

Lemma S2. Denote $\eta(x) = p^1(x) - p^0(x)$. Then given Assumption S3, $\eta(x)$ is Lipschitz continuous with constant $2L$.

Proof:

$$
\begin{aligned}
|\eta(x) - \eta(x')| &= |p^1(x) - p^0(x) - p^1(x') + p^0(x')| \\
&= |p^1(x) - p^1(x') - (p^0(x) - p^0(x'))| \\
&\leq |p^1(x) - p^1(x')| + |(p^0(x) - p^0(x'))| \\
&\leq 2L|x - x'|.
\end{aligned}
$$

□

[*] Equal contribution [1] Department of Computer Science, Aalto University, Espoo, Finland [2] Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA. Correspondence to: Iiris Sundin <iiris.sundin@aalto.fi>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

Definition: Let r be the smallest radius $r' > 0$ of a neighborhood $B_{r'}(x_e)$ of $x_e \in \Omega$, such that $|\eta(x)| = 0$ for some x in the border $\partial B_r(x_e)$.

Lemma S3. Given assumptions S3 and S4, then $r \geq \frac{h}{2L}$ for all $x_e \in \Omega$.

Proof: Counter-example: Show that if $r < \frac{h}{2L}$, then there does not exist $x \in B_r(x_e)$ for which $|\eta(x)| = 0$.

For any $x \in B_r(x_e)$ it holds that

$$\begin{aligned} |\eta(x_e) - \eta(x)| &\geq ||\eta(x_e)| - |\eta(x)|| \\ &= |\eta(x_e)| - |\eta(x)| \qquad (4) \\ \Leftrightarrow |\eta(x)| &\geq |\eta(x_e)| - |\eta(x_e) - \eta(x)| \\ &\geq h - 2L|x_e - x| \qquad (5) \\ &\geq h - 2Lr \\ &> h - 2L\frac{h}{2L} = 0, \qquad (6) \end{aligned}$$

where the equality in (4) comes from the fact that a necessary condition for $|\eta(x)| = 0$ is that $|\eta(x)| < |\eta(x_e)|$, (5) is by Assumption S4, and (6) is due to the counter-assumption $r < \frac{h}{2L}$.

Therefore $|\eta(x)|$ cannot be zero in $\partial B_r(x_e)$ unless $r \geq \frac{h}{2L}$. □

Lemma S4. Given assumption S3, $x \in \mathbb{R}$, and assuming $p^a(x_e) > p^{1-a}(x_e)$, then the expected number of samples from the group $1 - a$ in $B_r(x_e)$ is upper-bounded by $\mathbb{E}[n_{1-a}] \leq (1 - p(a))N(P^a_{B_r(x_e)} - \frac{h^2}{2L})$.

Proof:

$$\begin{aligned} &\mathbb{E}[n_{1-a}] \\ &= (1 - p(a))N \int_{B_r(x_e)} p^{1-a}(x)dx \\ &= (1 - p(a))N \int_{B_r(x_e)} \left(p^a(x) - (p^a(x) - p^{1-a}(x))\right) dx \\ &= (1 - p(a))N \left(P^a_{B_r(x_e)} - \int_{B_r(x_e)} (p^a(x) - p^{1-a}(x))dx\right) \\ &= (1 - p(a))N \left(P^a_{B_r(x_e)} - \int_{B_r(x_e)} |\eta(x)|dx\right) \\ &\leq (1 - p(a))N \left(P^a_{B_r(x_e)} - \frac{h^2}{2L}\right), \end{aligned}$$

where we have used $\int_{B_r(x_e)} |\eta(x)|dx \geq \frac{h^2}{2L}$ when $x \in \mathbb{R}$. This comes from the fact that $|\eta(x_e)| \geq h$ (Assumption S4), and by definition $|\eta(x)| = 0$ for some $x \in \partial B_r(x_e)$. Thus the integral has its smallest value when $|\eta(x)|$ decreases from h as fast as possible, that is, by Lipschitz constant $2L$ (Lemma S2), s.t. $|\eta(x)| = 0 \forall x \in \partial B_r(x_e)$. In case $x \in \mathbb{R}$, this yields the integrated area to be a triangle with height h, width $2r$ and area $\frac{1}{2}h2r \geq h\frac{h}{2L} = \frac{h^2}{2L}$ (Lemma S3). □

Theorem 1. Let N be the sample size, and a the treatment with the higher number of observations in $B_r(x_e)$, and $x \in \mathbb{R}$. Then the expected probability of Type S error in $B_r(x_e)$ has lower bound
$p(\text{“Type S error”}) > 2K^2D^2C^{N(P^a_{B_r(x_e)} - (1-p(a))\frac{h^2}{2L})}$.

Theorem 1 shows that, with fixed r, N and $p(a)$, the larger the local imbalance (h) in $B_r(x_e)$, the higher Type S error rate in $B_r(x_e)$ is.

Proof of Theorem 1. The expected number of samples of group a in $B_r(x_e)$ is $\mathbb{E}[n_a] = p(a)NP^a_{B_r(x_e)}$. The expected Type S error over all samples of size N from the true distribution is proportional to $\mathbb{E}[C^{(n_1+n_0)}] \geq C^{(\mathbb{E}[n_1]+\mathbb{E}[n_0])}$ (Lemma S1 and Jensen's inequality).

From this and Lemma S4 it follows that the expected Type S error in $B_r(x_e)$ has lower bound

$$\begin{aligned} &p(\text{“Type S error”}) > 2K^2D^2C^{\mathbb{E}[n_a]+\mathbb{E}[n_{1-a}]} \\ &\quad \geq 2K^2D^2C^{\left(p(a)NP^a_{B_r(x_e)}+(1-p(a))N(P^a_{B_r(x_e)}-\frac{h^2}{2L})\right)} \\ &\quad \geq 2K^2D^2C^{N\left(P^a_{B_r(x_e)}-(1-p(a))\frac{h^2}{2L})\right)}. \end{aligned}$$

□

In higher dimension, the key difference is in the result of Lemma S4, affecting the term $\frac{h^2}{2L}$. Specifically, the integral $\int_{B_r(x_e)} |\eta(x)|dx \geq M$, where M depends on the dimensionality of x; in one dimension $M = \frac{h^2}{2L}$ as in Lemma S4.

Now, we have shown that imbalance increases locally the Type S error rate. Then the question remains whether the error rate increases globally as well, or do the local effects cancel out each other. We prove this in one-dimensional case, but we see no reason why the proof would not extend to higher dimensions as well. The following assumption and theorem give conditions under which imbalance increases the global Type S error rate.

Assumption S5. Assume the following balanced and imbalanced settings. In the balanced setting, let $p^1(x) = p^0(x) = p(x)$, and $x \in \mathbb{R}$. Without loss of generality we assume that imbalance arises from a shift in $p^0(x)$, s.t. in the imbalanced setting $p^0(x) = p^1(x) - \eta(x)$, where $\eta(x) \in \mathbb{R}$, and $\int \eta(x)dx = 0$.

Lemma S5. In the imbalanced setting and under assumption S5, $p(x) = up^1(x) + (1-u)p^0(x) = p^1(x) - (1-u)\eta(x)$, where $u := p(a=1)$.

Proof. By simply: $p(x) = up^1(x) + (1-u)p^0(x) = up^1(x)+(1-u)(p^1(x)-\eta(x)) = up^1(x)+(1-u)p^1(x)-(1-u)\eta(x) = p^1(x) - (1-u)\eta(x)$.

Lemma S6. Given assumption S3, the maximum probability density at $x \in \mathcal{X}$ is $p_{\text{max}} \leq \sqrt{L}$.

Proof: By Assumption S3,

$$\begin{aligned}
&|p^1(x) - p^1(x')| \leq L|x - x'|, \quad \text{and} \\
&|p^0(x) - p^0(x')| \leq L|x - x'|, \quad \text{and} \\
&p(x) = up^1(x) {+} (1-u)p^0(x), \\
\Rightarrow &|p(x) - p(x')| \leq uL|x - x'| + (1-u)L|x - x'| \\
&\qquad = L|x - x'|.
\end{aligned}$$

Because $p(x)$ integrates to one, the highest possible density is achieved by first increasing $p(x)$ as quickly as possible to p_{max}, and then decreasing it back to zero; Otherwise some of the density would be spread to a wider range. Therefore, we get the maximum p_{max} by the sum of two triangles with height p_{max} and width $\frac{p_{\text{max}}}{L}$:

$$\begin{aligned}
&2 * \frac{1}{2} p_{\text{max}} (\frac{p_{\text{max}}}{L}) \leq 1 \\
\Leftrightarrow &p_{\text{max}} \leq \sqrt{L}.
\end{aligned}$$

□

The following theorem gives a sufficient condition for the increase of the expected global Type S error rate.

Theorem 2. Denote $P_{\eta \geq h} = \int_{\mathcal{X}} \mathbb{I}(\eta(x) \geq h) p_t(x) dx$, where $p_t(x)$ is the covariate distribution in the test set. Given Assumption S5, if $P_{\eta \geq h} > C^{N(1-u)h}$, then imbalance $\eta(x)$ increases the lower bound of the expected global Type S error rate in $\mathcal{X}$.

Proof of Theorem 2. We prove this in one-dimensional setting. The intuition is that since the error rate increases exponentially with decreasing number of samples, then in high-imbalance areas, where $\eta(x) \geq h$, the local increase in the error rate cannot be compensated elsewhere. We start by decomposing the lower bound to the bound without imbalance and a term that depends on imbalance. We then show that the imbalance-related term is greater than zero when $P_{\eta \geq h}$ is high enough, and therefore the imbalance increases the lower bound of the global Type S error rate.

In an infinitesimally small interval dx, the expected number of observations *over all samples* of size N from the true distribution, is $\rho_1 dx = \mathbb{E}[n_1] = uNp^1(x)dx$ and $\rho_0 dx = \mathbb{E}[n_0] = (1-u)Np^0(x)dx$.

Then, by Assumption S5, $\rho_1 + \rho_0 = uNp^1(x) + (1-u)N(p^1(x) - \eta(x)) = Np^1(x) - (1-u)N\eta(x) = N(p^1(x) - (1-u)\eta(x)) = Np(x)$. (Last equation from Lemma S5).

Then the expected effect on the expected Type S error is proportional to $\mathbb{E}[C^{(n_1+n_0)}] \geq C^{(\mathbb{E}[n_1]+\mathbb{E}[n_0])} = C^{(\rho_1+\rho_0)}$ (Jensen's inequality and Lemma S1), and the expected error rate in $\mathcal{X}$ is

$$\begin{aligned}
\gamma &\geq 2K^2D^2 \int_{\mathcal{X}} C^{(\rho_1+\rho_0)} p_t(x) dx \\
&= 2K^2D^2 \int_{\mathcal{X}} C^{Np(x)} p_t(x) dx.
\end{aligned}$$

Denote the expected error rate in the balanced setting as $\gamma_0 \geq 2K^2D^2 \int_{\mathcal{X}} C^{Np^1(x)} p_t(x) dx := b_0$, which comes from the Assumption S5. Then the expected error rate in the imbalanced setting has a lower bound

$$\gamma \geq 2K^2D^2 \int_{\mathcal{X}} C^{Np(x)} p_t(x) dx$$

which by Lemma S5 is

$$\begin{aligned}
&= 2K^2D^2 \int_{\mathcal{X}} C^{N(p^1(x)-(1-u)\eta(x))} p_t(x) dx \\
&= 2K^2D^2 \int_{\mathcal{X}} \Big(C^{Np^1(x)} C^{-N(1-u)\eta(x)} \\
&\qquad\qquad - C^{Np^1(x)} + C^{Np^1(x)}\Big) p_t(x) dx \\
&= 2K^2D^2 \int_{\mathcal{X}} C^{Np^1(x)} \left(C^{-N(1-u)\eta(x)} - 1\right) p_t(x) dx \\
&\qquad\qquad + 2K^2D^2 \int_{\mathcal{X}} C^{Np^1(x)} p_t(x) dx \\
&= 2K^2D^2 \int_{\mathcal{X}} C^{Np^1(x)} \left(C^{-N(1-u)\eta(x)} - 1\right) p_t(x) dx + b_0 \\
&\geq 2K^2D^2 \int_{\mathcal{X}} C^{Np_{\text{max}}} \left(C^{-N(1-u)\eta(x)} - 1\right) p_t(x) dx + b_0,
\end{aligned}$$

and by Lemma S6

$$\geq 2K^2D^2 C^{N\sqrt{L}} \left(\int_{\mathcal{X}} C^{-N(1-u)\eta(x)} p_t(x) dx - 1 \right) + b_0.$$

Since b_0 is the lower bound in the balanced setting, the lower bound of the expected Type S error rate increases with increasing imbalance, if $\int_{\mathcal{X}} C^{-N(1-u)\eta(x)} p_t(x) dx > 1$.

Next, we consider when does this condition hold. Denote the set where $\eta(x) \geq 0$ as $\mathcal{X}^+$, and similarly $\mathcal{X}^-$ the set where $\eta(x) < 0$. Then

$$
\begin{aligned}
&\int_{\mathcal{X}} C^{-N(1-u)\eta(x)} p_t(x)dx \\
&= \int_{\mathcal{X}+} C^{-N(1-u)|\eta(x)|} p_t(x)dx \\
&+ \int_{\mathcal{X}-} C^{N(1-u)|\eta(x)|} p_t(x)dx \\
&\geq \int_{\mathcal{X}+\setminus\mathcal{X}_{\eta\geq h}} C^{-N(1-u)|\eta(x)|} p_t(x)dx \\
&+ \int_{\mathcal{X}_{\eta\geq h}} C^{-N(1-u)h} p_t(x)dx + \int_{\mathcal{X}-} C^{N(1-u)\eta_{\text{max}}} p_t(x)dx \\
&\geq \int_{\mathcal{X}+\setminus\mathcal{X}_{\eta\geq h}} p_t(x)dx \ + \ C^{-N(1-u)h} \int_{\mathcal{X}_{\eta\geq h}} p_t(x)dx \\
&+ C^{N(1-u)\eta_{\text{max}}} \int_{\mathcal{X}-} p_t(x)dx \\
&= P_{0\leq\eta<h} \ + \ C^{-N(1-u)h} P_{\eta\geq h} \ + \ C^{N(1-u)\eta_{\text{max}}} P_{\eta<0} \\
&\geq C^{-N(1-u)h} P_{\eta\geq h} \\
&> 1, \quad \text{if } \ P_{\eta\geq h} > C^{N(1-u)h}.
\end{aligned}
$$

Here η_{max} is the maximum difference between the the distributions $p^1(x)$ and $p^0(x)$, and
$\mathcal{X}_{\eta\geq h} = \{x \in \mathcal{X} \mid \eta(x) \geq h\}$, $h > 0$. □

2. Details of the Implementations

2.1. The Observed and Estimated Type S Error Rate in Imbalanced Data

Data is generated from

$$
\begin{aligned}
x &\sim N(0,1) \\
a &\sim \text{Bernoulli}(\theta_x) \\
b_0, b_1 &\sim N(0, 0.5) \\
y \mid x, a &\sim N(f(x) + (\beta_0 + \beta_1 x)a, \sigma_0^2), \ \text{and} \\
f(x) &= 2\left(\frac{1}{1+e^{-x+b}} - 0.5\right),
\end{aligned}
$$

where imbalance is generated by setting $\theta_x = e_x$ for $x \leq 0$ and $\theta_x = 1 - e_x$ for $x > 0$. Here $e_x = p(a = 1 \mid x)$ is the propensity score. Technical details: The shape of $f(x) \in (-1, 1)$ is chosen to be half of a sigmoid within range of 1σ from $\bar{x}$, so as to either have a saturating effect or an increasingly increasing effect (defined by the sign of $b \in \{-1, 1\}$, $b \sim$ uniform).

The outcome model generation (b_0, b_1 and b) is repeated 200 times, and for each outcome model we generate 6 training sets. The training data generation differs in the propensity scores $e_x \in \{0.0, 0.1, ..., 0.5\}$, resulting in different levels of imbalance in the training data sets. The size of the test set is 500.

We measure imbalance using the Maximum Mean Discrepancy (MDD) (Gretton et al., 2012), with Gaussian kernel and length-scale 0.8. We model the potential outcomes using two independent Gaussian Processes with squared exponential kernel.

2.2. Simulated Example

Synthetic data. The outcome $y \in \{0, 1\}$ is Bernoulli distributed with parameter $\theta_{x,a}$, given a one-dimensional covariate $x \in \mathbb{R}$ and treatment $a \in \{0, 1\}$. The data is generated from a logistic regression model with interaction between a and 3 radial basis functions (RBF) $\phi(x)$. The data is generated as follows:

$$
\begin{aligned}
x &\sim \text{uniform}(-4.5, 4.5) \\
a \mid x &= 1 \ \text{ if } \ x < -1.5, \ \ 0 \ \text{ else} \\
y \mid x, a &\sim \text{Bernoulli}(\theta_{x,a}),
\end{aligned}
$$

where $\theta_{x,a} = \text{logit}^{-1}(\mathbf{w}_0^\top \phi(x) + \mathbf{w}_1^\top \phi(x)a)$, and RBF centers are at $-3, 0, 3$, have lenght-scale 1, and $\mathbf{w}_0 = [0.5\ 1.5\ 1.5]^\top$, $\mathbf{w}_1 = [1\ -1\ -3.0]^\top$.

Training sample size is 30. The 9 test points are set with equal distance to each other in the range of x. Data generation is repeated 100 times, and the reported values are the mean and the 95% bootstrap confidence intervals.

Model and learning. We model the data with a logistic regression model $p(y \mid x, a) \sim \text{Bernoulli}(\theta_{x,a})$, where $\theta_{x,a}$ has the same form as in the data generation process. The model is fit using a probabilistic programming language Stan (Stan Development Team, 2017; Carpenter et al., 2017). We assume that the RBF centers and length-scale are known ($-3, 0, 3$, and lenght-scale 1), so that only $\mathbf{w}_0$ and $\mathbf{w}_1$ need to be learned.

2.3. Gaussian Process Model with Direct Feedback

For modeling the Gaussian process with direct feedback on patient response, a Gaussian process prior with squared exponential covariance function and Gaussian likelihood was used. Responses for different treatments were modeled with independent models. We use Gamma distribution with shape 1.5 and rate 3.0 as prior for lengthscale, variance and noise. The models were implemented with GPy-framework [1]. Since the observed data and the counterfactual feedback were obtained from different sources, both were assumed to have separate noise priors. Since different attributes have very different effect on the response, the covariance function used different lengthscale parameters for different dimen-

[1] GPy v. 1.9.2 https://github.com/SheffieldML/GPy

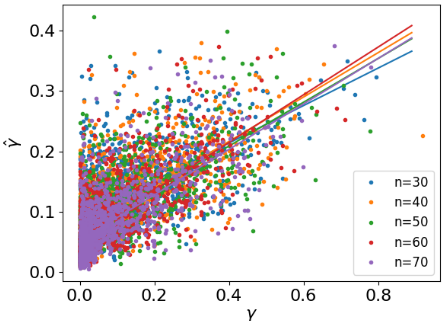

Figure 1. The observed (γ) and estimated ($\hat{\gamma}$) Type S error rate in 1200 data sets. Low estimated error indicates low observed error. Solid lines show the regression line for each sample size.

sions. Hyper-parameters were estimated by maximizing the marginal likelihood.

3. Additional results

3.1. The Observed and Estimated Type S Error Rate in Imbalanced Data

The observed and estimated Type S error rates from the experiment described in Section 2.1 (Section 5.1 in the paper) are shown in Figure 1. The results show that low estimated error indicates low observed error.

3.2. K-Nearest Neighbor Approximation of D-M aware

We additionally tested the idea to approximate full D-M aware by only computing the expected minimization of Type S error rate for the k nearest neighbors of the test unit. Important here is that we use the *model's distance measure*, which in the case of GPs is the optimized kernel (with Automatic Relevance Determination ARD). Our preliminary results (Fig. 2) show that selecting too few neighbors will impair the performance of active learning.

References

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017. URL https://www.jstatsoft.org/v076/i01.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085, 2017.

Stan Development Team. Pystan: the python interface to stan, version 2.16.0.0, 2017. URL http://mc-stan.org.

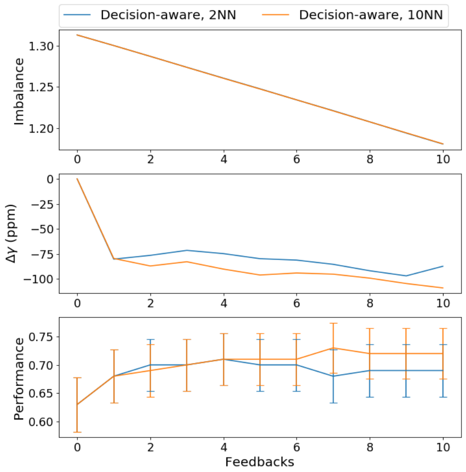

Figure 2. Comparison of the D-M aware active learning using k=2 (Decision-aware 2NN) and k=10 (Decision-aware 10NN) nearest neighbor approximations shows that the performance degrades if the number of neighbors k is too low.