**ROB313: Introduction to Learning from Data**
**University of Toronto Institute for Aerospace Studies**

# Assignment 5 (11 pts)

**Due April 13, 2019, 23:59**

**Q1) 7pts** Perform Bayesian inference for a logistic regression model with a Bernoulli likelihood (as we had considered in A3), and with a zero-centered, uncorrelated Gaussian prior on the weights, as follows:

$$\Pr(y|\mathbf{w},\mathbf{x}) = \big[\widehat{f}(\mathbf{x};\mathbf{w})\big]^{y}\big[1-\widehat{f}(\mathbf{x};\mathbf{w})\big]^{1-y},$$
$$\Pr(\mathbf{w}) = \prod_{i=0}^{D}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{w_i^2}{2\sigma^2}\Big) = \mathcal{N}\big(\mathbf{w}\big|\mathbf{0},\sigma^2\mathbf{I}\big),$$

where $\widehat{f}(\mathbf{x};\mathbf{w}) = \Pr(y{=}1|\mathbf{w},\mathbf{x})$ gives the class conditional probability of class 1 by mapping $\mathbb{R}^D \to [0,1]$, and $\mathbf{w} = \{w_0, w_1, \ldots, w_D\} \in \mathbb{R}^{D+1}$. Also, $\widehat{f}$ is a logistic sigmoid acting on a linear model as follows

$$\widehat{f}(\mathbf{x};\mathbf{w}) = \mathsf{sigmoid}\Big(w_0 + \sum_{i=1}^{D} w_i x_i\Big),$$

where $\mathsf{sigmoid}(z) = \frac{1}{1+\exp(-z)}$. Making the assumption that all training examples are *i.i.d.*, the log-likelihood, and log-prior, along with their gradient ($\nabla$) and hessian ($\nabla^2$) can be written as follows

$$\log\Pr(\mathbf{y}|\mathbf{w},\mathbf{X}) = \sum_{i=1}^{N} y^{(i)}\log\big(\widehat{f}(\mathbf{x}^{(i)};\mathbf{w})\big) + \big(1-y^{(i)}\big)\log\big(1-\widehat{f}(\mathbf{x}^{(i)};\mathbf{w})\big),$$
$$\nabla\log\Pr(\mathbf{y}|\mathbf{w},\mathbf{X}) = \sum_{i=1}^{N}\big[y^{(i)} - \widehat{f}(\mathbf{x}^{(i)};\mathbf{w})\big]\ \bar{\mathbf{x}}^{(i)},$$
$$\nabla^2\log\Pr(\mathbf{y}|\mathbf{w},\mathbf{X}) = \sum_{i=1}^{N}\widehat{f}(\mathbf{x}^{(i)};\mathbf{w})\ \big[\widehat{f}(\mathbf{x}^{(i)};\mathbf{w})-1\big]\ \bar{\mathbf{x}}^{(i)}\bar{\mathbf{x}}^{(i)T}$$
$$\log\Pr(\mathbf{w}) = -\frac{D+1}{2}\log(2\pi) - \frac{D+1}{2}\log(\sigma^2) - \sum_{i=0}^{D}\frac{w_i^2}{2\sigma^2},$$
$$\nabla\log\Pr(\mathbf{w}) = -\frac{\mathbf{w}}{\sigma^2}$$
$$\nabla^2\log\Pr(\mathbf{w}) = -\frac{1}{\sigma^2}\mathbf{I}$$

where $\bar{\mathbf{x}}^{(i)} = \big\{1, x_1^{(i)}, \ldots, x_D^{(i)}\big\}^T \in \mathbb{R}^{D+1}$. All studies will be done on the `iris` dataset, with the training and validation sets merged, and considering only the second response to determine whether the flower is an *iris virginica*, or not[1]. You are encouraged to re-use code from previous assignments where possible.

[1] Use `x_train, x_test = np.vstack((x_train, x_valid)), x_test` and `y_train, y_test = np.vstack((y_train[:,(1,)], y_valid[:,(1,)])), y_test[:,(1,)]`

(a) **(2pts)** Consider the prior variances $\sigma^2 = 0.5$, $\sigma^2 = 1$, and $\sigma^2 = 2$. Which of these priors gives a model with a higher complexity? Explain why. Choose between these prior variances by approximating the log marginal likelihood using a Laplace approximation. Report your results.

(b) **(2.5pts)** Choose a proposal distribution and use importance sampling to estimate the most probable predictive posterior class on each element of the test set using a prior variance of $\sigma^2 = 1$. Report your test set accuracy results and justify your chosen proposal. Also, analyze and comment on the accuracy of your proposal distribution. It may help to visualize the values of the posterior evaluated at your samples to help justify your answer. You may find the `scipy.stats` module useful for sampling.

(c) **(2.5pts)** Write a Metropolis-Hastings MCMC sampler to estimate the most probable predictive posterior class on each element of the test set using a prior variance of $\sigma^2 = 1$. Use the proposal $q(\mathbf{w}^*|\mathbf{w}^{(i)}) = \mathcal{N}\big(\mathbf{w}^*\big|\mathbf{w}^{(i)}, \sigma_p^2\mathbf{I}\big)$, and burn-in 1000 iterations. After the burn-in, sample for an additional 10000 iterations, thinning by collecting every 100th sample. Choose and report your value of $\sigma_p^2$, and report your test set accuracy results. Also, plot the predictive posterior class-conditional probability samples (i.e. plot $\Pr(y^{(*)}{=}1|\mathbf{x}^{(*)}, \mathbf{w}^{(i)})$, where $\mathbf{w}^{(i)} \sim \Pr(\mathbf{w}|\mathbf{X}, \mathbf{y})$ for $i = 1, \ldots, S$) for the 9th, and the 10th flowers in the testing dataset[2] as a histogram. Discuss the visualization. What valuable information from these figures would be lost if a frequentist approach were taken?

**Q2) 4pts** Read, and write a brief report on one of the research papers listed below. These papers cover a range of "meta" topics that are relevant to engineers in machine learning. You may select any one of the papers that piques your interest. Your report should

- start off with a high level description of the paper;
- summarize the paper as if you were explaining the concept to a classmate[3];
- include your personal thoughts about strengths and weaknesses of the material;
- not repeat text from the paper or the abstract.

Your response should be approximately half a page. Do not exceed one page. We encourage you to post your summary on https://www.shortscience.org/

**Safety in Machine Learning Papers**

Varshney, Kush R. *Engineering safety in machine learning.* Information Theory and Applications Workshop (ITA), IEEE, 2016.
https://arxiv.org/pdf/1601.04126.pdf

Amodei, Dario, et al. *Concrete problems in AI safety.* arXiv:1606.06565, 2016.
https://arxiv.org/pdf/1606.06565.pdf

[2] I.e. `x_test[[9,10]]` and `y_test[[9,10]]`

[3] Alternate instructions are included for the last paper on the list, for which a summary is not to be included in your report.

**Fairness in Machine Learning Papers**

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. *Fairness constraints: Mechanisms for fair classification.* In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 962–970, 2017.
https://arxiv.org/pdf/1507.05259.pdf

**Practical Aspects of Machine Learning Papers**

Domingos, P. *A few useful things to know about machine learning.* Communications of the ACM, 55(10), pp. 78–87, 2017.
https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf
*This paper is a bit different than the others in that it is already a summary of many topics so it would not make sense for you to provide a summary yourself. Instead, write a critical review of the paper: What did you like? What did you disagree with? What was missing in the paper?*

**Submission guidelines:** Submit an **electronic copy** of your report in **pdf** format, and **documented** python scripts. You should include a file named "README" outlining how the scripts should be run. Upload a single tar or zip file containing all files to Quercus. You are expected to verify the integrity of your tar/zip file before uploading. Do not include (or modify) the supplied *.npz data files or the data_utils.py module in your submission. The report must contain

- Objectives of the assignment
- A brief description of the structure of your code, and strategies employed
- Relevant figures, tables, and discussion

Do not use scikit-learn for this assignment, the intention is that you implement the simple algorithms required from scratch. Also, for reproducibility, always set a seed for any random number generator used in your code. For example, you can set the seed in numpy using `numpy.random.seed`