COMP 6321, Winter 2014 **CONCORDIA UNIVERSITY**

MACHINE LEARNING.
Assignment 2.
Due: Feb. 14, 2014.

1. [30 points] **L1 vs. L2 Regularization**

   In this problem, we explore the differences between L1 and L2 regularization, discussed in lecture 2. You are given two data files, hw2x.dat and hw2y.dat (containing inputs and outputs, respectively).

   (a) [10 points] Split the data into a training set, containing 90% of the instances, and a test set, containing 10% of the instances (since this is exploratory, we will not do full cross-validation). Write Matlab code to perform L2 regularization. Plot on one graph the root mean squared error on the training set, and the root mean squared error on the test set, as a function of the regularization parameter λ. Vary λ starting at 0, and go high enough that you can see an "interesting" range of behavior. Plot, on a different graph, all of the weights, as a function of λ

   (b) [10 points] Using the quadprog function of Matlab, write a function that performs L1 regularization

   (c) [10 points] Plot the same graphs as above for the L1 regularization. Explain what you observe, and comment on how you think the data was generated.

2. [10 points] **Dealing with missing data**

   Suppose that you use a Gaussian discriminant classifier, in which you model explicitly $P(y = 1)$ (using a binomial) and $P(\mathbf{x}|y = 0)$ and $P(\mathbf{x}|y = 1)$. The latter have distinct means μ_0 and μ_1, and a shared covariance matrix Σ (a frequent assumption in practice). Suppose that you are asked to classify an example for which you know inputs $x_1, \ldots x_{n-1}$, but the value of x_n is missing. In practice, a common approach in this case is to "fill in" the value of x_n by its class-conditional means, $E(x_n|y = 0)$ and $E(x_n|y = 1)$. Using the log-odds ratio, give a mathematical justification for this approach.

3. [10 points] **Naive Bayes assumption**

   We discussed in class the fact that naive Bayes assumes that the inputs are conditionally independent given the class label. Suppose now that we have a problem with two binary inputs, x_1 and x_2, which are truly conditionally independent given the class label. Suppose that, by accident, we add a feature, x_3, which is an identical copy of x_2.

   (a) [2 points] How many parameters does the initial Naive Bayes classifier have? How many parameters does the classifier with three features have?

(b) [8 points] Assuming that the parameters are learned perfectly, describe (using formulas) the effect of the added duplicate feature on the decision boundary of the naive Bayes classifier (by considering the log-odds ratio). What is the worst-case scenario? Explain in 1-2 sentences what this means for the robustness of the naive Bayes classifier.

4. [10 points] **MAP estimates vs. ML estimates**

We discussed in class briefly the idea of using Bayesian priors for regularization. Consider using the logistic regression model, in which (as discussed) we find a parameter vector $\mathbf{w}$ which maximizes the conditional probability of the class labels, given the input:

$$\arg\max_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i;\mathbf{w})$$

Let us call this estimates $\mathbf{w}_{ML}$. Now suppose that we put a Bayesian prior over the parameters of the form: $\mathbf{w} \sim \mathcal{N}(0, \tau^2 I)$, where $\tau > 0$ and I is the identity matrix. Now we will find the parameters as:

$$\mathbf{w}_{MAP} = \arg\max_{\mathbf{w}} \prod_{i=1}^{m} P(y_i|\mathbf{x}_i;\mathbf{w})P(\mathbf{w})$$

What is the relationship between $\|\mathbf{w}_{MAP}\|_2$ and $\|\mathbf{w}_{ML}\|_2$?

5. [40 points] **Using discriminative vs. generative classifiers**

For this problem, you will experiment with a version of the Wisconsin data set that we use as an illustration in class. The data is available in files `wpbcx.dat` and `wpbcy.dat`.

(a) [10 points] Implement logistic regression. If you use a learning-rate version, you will need to set up your code in such a way as to be able to search for a good learning rate. You can also use the iterative recursive least-squares version (whichever you prefer).

(b) [10 points] In a first experiment, use just a bias term and the first feature (first column in the wpbcx.dat file). Set up a Gaussian naive Bayes classifier, and compare its results with logistic regression, using 10-fold cross-validation. Comment on what you observe.

(c) [10 points] Implement Gaussian naive Bayes for using all the inputs.

(d) [10 points] Compare the performance of Gaussian naive Bayes and logistic regression, using 10-fold cross-validation. Write a little write-up describing *precisely* your empirical setup (including any parameter choices), your observed results, and the conclusions you draw from them. Note that "precise" means that someone reading your description could re-implement your experiments and would get results consistent with yours (this is the standard you should use for experimental descriptions in papers as well).

6. [15 points] (BONUS) **Conjugate priors**

   (a) [10 points] (Bishop, PRML, Problem 2.44) Consider a univariate Gaussian distribution $N(x|\mu, \tau^{-1})$ having conjugate Gaussian-gamma prior given by (2.154), i.e.,

   $$p(\mu, \lambda) = N(\mu|\mu_0, (\beta\lambda)^{-1})Gam(\lambda|a, b)$$

   and a data set $\mathbf{x} = \{x_1, \ldots, x_N\}$ of i.i.d. observations. Show that the posterior distribution is also a Gaussian-gamma distribution of the same functional form as the prior, and write down expressions for the parameters of this posterior distribution.

   (b) [5 points] (Bishop, PRML, Problem 2.45) Verify that the Wishart distribution defined by (2.155), i.e.,

   $$W(\mathbf{\Lambda}|\mathbf{W}, \nu) = B|\mathbf{\Lambda}|^{(\nu-D-1)/2} \exp\left(-\frac{1}{2}Tr(\mathbf{W}^{-1}\mathbf{\Lambda})\right)$$

   is indeed a conjugate prior for the precision matrix of a multivariate Gaussian.