# Statistical Learning-Classification
# STAT 441 / 841, CM 763

Assignment 4
Department of Statistics and Actuarial Science
University of Waterloo

| |
|---|
| **Due: Monday Dec 3, at 1 pm** |
| **Policy on Lateness:** Late assignments are NOT accepted. |

1. In a binary classification problem where $y \in \{0, 1\}$

   a) Assume $\mathbf{x} = (x_1, \ldots, x_d)^T$, and $x_j \in \{0, 1\}$, for $j = 1 \ldots d$. Define $P(y = 1) = p$, $P(x_j = 1|y = 1) = p_{j1}$, and $P(x_j = 1|y = 0) = p_{j0}$. Show that the Naive Bayes classifier is equivalent to a linear classification rule in the form of $\hat{y} = sign(\mathbf{w}.\mathbf{x} - b)$. Write $\mathbf{w}$ and $b$ in terms of $p, p_{i1}$, and $p_{i0}$.

   b) Now suppose $x_j \in \mathbb{R}$. Assume $P(y = 1) = p$, $x_j|y = 1 \sim N(\mu_{j1}, \sigma_j^2)$, and $x_j|y = 0 \sim N(\mu_{j0}, \sigma_j^2)$. Show that the Naive Bayes classifier is equivalent to a linear classification rule in the form of $\hat{y} = sign(\mathbf{w}.\mathbf{x} - b)$. Write $\mathbf{w}$ and $b$ in term of $p, \mu_{j1}, \mu_{j0}$, and $\sigma_j$.

2. Suppose $X_{1:}, \ldots X_{10:}$ are standard independent Gaussian, and the target $y$ is defined by

$$y = \begin{cases} 1 & if \ \sum_j^{10} X_{j:}^2 > 9.34 \\ -1 & otherwise \end{cases}$$

   Sample 2000 training cases, with approximately 1000 points in each class, and 10,000 test observations.

   a) Write a program implementing AdaBoost with stumps.

   b) Plot the training error as well as test error, and discuss its behavior.

   c) Investigate the number of iterations needed to make the test error finally start to rise.

3. In the maximum-margin hyperplane problem, let's $\tau$ denotes the value of the margin. Show that

$$\frac{1}{\tau^2} = 2\sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

where $k(\mathbf{x}_i, \mathbf{x}_j)$ is a valid kernel.

**Only for Grad Students**

4. Kernel functions can be defined over objects as diverse as graphs, sets, and text documents. For instance consider the space of all possible subsets $A$ of a given fixed set $D$. Show that the kernel function $k(A_1, A_2) = 2^{|A_1 \cap A_2|}$ corresponds to an inner product in a feature space of dimensionality $2^{|D|}$ defined by the mapping $\phi(A)$. Here $A$ is a subset of $D$ , $A_1 \cap A_2$ denotes the intersection of sets $A_1$ and $A_2$, and $|A|$ denotes the number of elements in $A$. Find mapping $\phi(A)$ such that $k(A_1, A_2) = \phi(A_1)^T \phi(A_2)$.

5. We learned the effect of $L_2$ norm in class. To show the effect of $L_2$ norm, we started by a quadratic approximation to the objective function.

$$\hat{J}(\theta) = J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w^*)$$

We used singular value decomposition of $H = Q\Lambda Q^T$ and concluded that the effect of weight decay is to rescale the coefficients of eigenvectors.

$$\tilde{w} = Q(\Lambda + \alpha I)^{-1} \Lambda Q^T w^*$$

The $i$th component is rescaled by a factor of $\frac{\lambda_i}{\lambda_i + \alpha}$.

Do a similar analysis for *Early Stopping*, and show its effect.

Early Stopping can be used as a form of Regularization. In Early Stopping, Instead of running our optimization algorithm until we reach a (local) minimum of validation error, we run it until the error on the validation set has not improved for some amount of time.

Similar to $L_2$ norm analysis, start by taking a quadratic approximation to the objective function $J$ in the neighborhood of the empirically optimal value of the weights $w^*$.

$$\hat{J}(\theta) = J(w^*) + \frac{1}{2}(w - w^*)^T H (w - w*)$$

Consider updating the parameters via gradient descent.

$$w^{(\tau)} = w^{(\tau-1)} - \eta \nabla_w J(w^{(\tau-1)})$$

Show that after $\tau$ training updates,

$$w^{(\tau)} = Q[I - (I - \eta\Lambda)^\tau]Q^T w^*.$$

Assume $w^{(0)} = 0$, and $|1 - \eta\lambda_i| < 1$, where $\eta$ is the learning rate.