CO902
Problem Set 4

1. *Estimating "a priori" probabilities. This is Bishop Ex 4.9 in my own notation; see B §4.2.2 for hints.* Consider a generative classification model for K classes defined by prior class probabilities $P(Y = k) = \pi_k$ and class-conditional densities $p_k(\mathbf{X}) = p(\mathbf{X}|Y = k)$, where $\mathbf{X} \in R^d$ is the input feature vector and $Y \in \{1, 2, \ldots, K\}$ is the true class[1]. For data $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, N$, considering the joint likelihood for $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and the (unspecified) parameters of the data $\mathbf{x}$; find the maximum likelihood estimate of the class frequencies $\boldsymbol{\pi}$. Hint: Express the class variable as $t_{ik} = \delta_{y_i k}$, where $\delta_{y_i k}$ is 1 if $y_i = k$ and 0 otherwise, so that $\mathbf{t}_i$ is a K-vector with $K - 1$ zeros and 1 one, indicating the true class for observation i.

2. *Optimal decision rule for continous data.* Consider supervised learning based on $\{\mathbf{X}_i, Y_i\}$, $i = 1, \ldots, n$, for data $\mathbf{X}_i \in \Re^d$ and a class membership $Y_i \in \{1, 2, \ldots, K\}$. Show that the optimal decision rule $D_\mathbf{x}$ takes the form

$$D_\mathbf{x} = \operatorname*{argmax}_k P(Y = k|\mathbf{X} = \mathbf{x})$$

3. *Regression.* Consider observations of the form $\{\mathbf{X}_i, Y_i\}$, $i = 1, \ldots, n$, for predictors $\mathbf{X}_i \in \Re^d$ and response $Y_i \in \Re$. Let $\mathbf{X}$ be a $n \times (d + 1)$ matrix, where each row consists of $[\,1\ \mathbf{X}_i^\top\,]$, and $\mathbf{Y}$ be the n-vector of responses. The linear regression model approximates $\mathbf{Y}$ with

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{w}$$

where $\mathbf{w}$ is a $d + 1$ vector of regression coefficients. The standard estimate of $\mathbf{w}$ is $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$.

   (a) Derive $\hat{\mathbf{w}}$ as the minimizer of the residual sum of squares,

$$J(\mathbf{w}) = (\mathbf{Y} - \mathbf{X}\mathbf{w})^\top(\mathbf{Y} - \mathbf{X}\mathbf{w}).$$

   (b) Derive $\hat{\mathbf{w}}$ on the assumption that $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\mathbf{w}, \mathbf{I}_n\sigma^2)$, where $\mathcal{N}_n$ is a n-dimensional multivariate Normal distribution, $\mathbf{I}_n$ is a $n \times n$ identity matrix, and σ^2 is the resdiual error variance.

4. *Ridge Regression.* Consider the same data matrix $\mathbf{X}$ and response $\mathbf{Y}$ as in the previous question.

   (a) The following cost function is the residual sum of squares penalized by the sum of squares of the regression coefficients,

$$J(\mathbf{w}) = (\mathbf{Y} - \mathbf{X}\mathbf{w})^\top(\mathbf{Y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w}^\top\mathbf{w}.$$

[1]Here, capital Roman letters indicate (yet to be observed) random variables, while lower case Roman letters indicate particular (observed) values of the random variables. Boldface font indicates a vector-valued variable.

Show that the ridge regression estimator $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{d+1})^{-1}\mathbf{X}^\top\mathbf{Y}$ minimizes this $J(\mathbf{w})$.

(b) Derive the ridge regression estimator as the maximum a postiori (MAP) estimator of a Bayesian model with prior

$$\mathbf{w} \sim \mathcal{N}_{d+1}(\mathbf{0}, \mathbf{I}_{d+1}\sigma_0^2),$$

where $\mathbf{I}_{d+1}$ is the $(d+1) \times (d+1)$ identity matrix and σ_0^2 is the prior variance. Explain the relationship beteween λ, σ and σ_0?