EE621 ADAPTATION AND LEARNING **Instructor: Ali H. Sayed**
Homework Assignment #3 (Bayesian Inference) **Due: April 29, 2019**

1) A random variable $\boldsymbol{y}$ follows a binomial distribution with parameters N and p, i.e., the probability of observing k successes in N trials is given by:

$$\mathbb{P}(\boldsymbol{y}=k) = \binom{N}{k} p^k (1-p)^{N-k}, \quad k = 0, 1, \ldots, N$$

Having observed $\boldsymbol{y} = y$, we wish to estimate the probability of success, p, using a MAP estimator. To do so, we assume that the marginal distribution of $\boldsymbol{p}$ follows a beta distribution with parameters (a, b), namely,

$$f_{\boldsymbol{p}}(p; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, p^{a-1}(1-p)^{b-1}, \quad 0 \le p \le 1$$

Show that

$$\widehat{p}_{\text{MAP}} \;=\; \frac{y+a-1}{N+a+b-2}$$

and compare with the ML solution for this case.

2) **True or False**:
   (a) The LDA classifier is a Bayes classifier.
   (b) The minimum distance classifier is a Bayes classifier.
   (c) The minimum distance classifier is a linear classifier.
   (d) The Fisher classifier is a Bayes classifier.
   (e) The Fisher classifier has a discriminant function structure.

3) Refer to the derivation of the PCA procedure. Consider an $M \times M$ symmetric and positive-definite matrix $\widehat{R}_p$. We seek to approximate it by a symmetric and positive-definite matrix $\overline{R}_p$ of size $M' \times M'$, where $M' < M$, by solving $\min_{\overline{R}_p} \|\widehat{R}_p - \overline{R}_p\|_F^2$, in terms of the Frobenius norm. Determine $\overline{R}_p$.

4) Consider the ℓ_2-regularized logistic risk

$$R(w) \;\stackrel{\Delta}{=}\; \rho\|w\|^2 \;+\; \mathbb{E}\left\{\ln\left(1+e^{-\boldsymbol{\gamma}\widehat{\boldsymbol{\gamma}}(w)}\right)\right\}, \quad \widehat{\boldsymbol{\gamma}}(w) = \boldsymbol{h}^{\mathsf{T}} w$$

and denote its minimizer by w^o. Prove that
   (a) $\|w^o\| \le \mathbb{E}\,\|\boldsymbol{h}\|/2\rho$.
   (b) $\|w^o\|^2 \le \text{Tr}(R_h)/4\rho^2$, where $\boldsymbol{h}$ is zero-mean and $R_h = \mathbb{E}\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{T}}$.

5) Show that the Perceptron algorithm can learn to implement a NAND function? Consider the other logical operations represented by NOT, AND, OR, NOR, XOR (exclusive OR), and XNOR (exclusive NOR). Show how each of these logical operations can be implemented by using solely NAND gates.

6) Consider two feature vectors $\{+1, h_a\}$ and $\{-1, h_b\}$, where h_a belongs to class $+1$ and h_b belongs to class -1. Assume that these two vectors meet the margin in an SVM implementation, that is, they satisfy $h_a^{\mathsf{T}} w - \theta = +1$ and $h_b^{\mathsf{T}} w - \theta = -1$. The parameters (w, θ) describe the separating hyperplane that lies in the center of the boundary region with maximal margin value. Project the vector difference $h_a - h_b$ along the unit-norm normal to the separating hyperplane and determine the size of the margin, $m(w)$, associated with w.

7) Refer to the case in which the entries of a feature vector $\boldsymbol{h}$ follow a multinomial distribution. Show that the naïve Bayes classifier reduces to solving $\max_{1 \le r \le R}\; \log(\pi_r) + h^{\mathsf{T}} w_r$, where $w_r = \begin{bmatrix} \log(p_{r1}) & \log(p_{r2}) & \ldots & \log(p_{rM}) \end{bmatrix}$.

8) Consider a cluster $\mathcal{C}$, of cardinality $|\mathcal{C}|$, and consisting of $M-$dimensional feature vectors, denoted generically by $h \in \mathcal{C}$. Let μ denote the mean of the cluster, i.e., the mean of the vectors in $\mathcal{C}$. For any vector $x \in \mathbb{R}^M$, show that

$$\sum_{h \in \mathcal{C}} \|h - x\|^2 = \sum_{h \in \mathcal{C}} \|h - \mu\|^2 \;+\; |\mathcal{C}|\, \|\mu - x\|^2$$

Computer project. Select a dataset of your choice from the repository https://archive.ics.uci.edu/ml/datasets.html. Reduce the feature dimension down to two using PCA and simulate the classification performance of Perceptron, soft-SVM, and logistic regression. Generate representative plots. Perform cross-validation to select the regularization parameter ρ for soft-SVM.