Machine Learning (CS 6140)

Sample Midterm Questions

Instructor: Ehsan Elhamifar

1) Show that the Euclidean distance from a point $\boldsymbol{x}$ to the hyperplane $\boldsymbol{w}^\top \boldsymbol{x} = 0$ is given by $\frac{|\boldsymbol{w}^\top \boldsymbol{x}|}{\|\boldsymbol{w}\|_2}$.

2) Consider the problem of separating data $\mathcal{D} = \{(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^N, y^N)\}$ from two classes with labels $\{-1, +1\}$, using the hyperplane $\boldsymbol{w}^\top \boldsymbol{x} = 0$. a) Derive an optimization on $\boldsymbol{w}$ in order to find the maximum geometric margin hyperplane. b) Write down the Lagrangian of the optimization. c) Derive the dual optimization.

3) We consider here a discriminative approach for solving the classification problem illustrated in Figure 1. We attempt to solve the binary classification task depicted in the Figure 1 with the simple linear logistic regression model

$$P(y = 1 | \, \boldsymbol{x}, \, \boldsymbol{w}) = g(w_0 + w_1 x_1 + w_2 x_2) = \frac{1}{1 + \exp\left(-w_0 - w_1 x_1 - w_2 x_2\right)}.$$

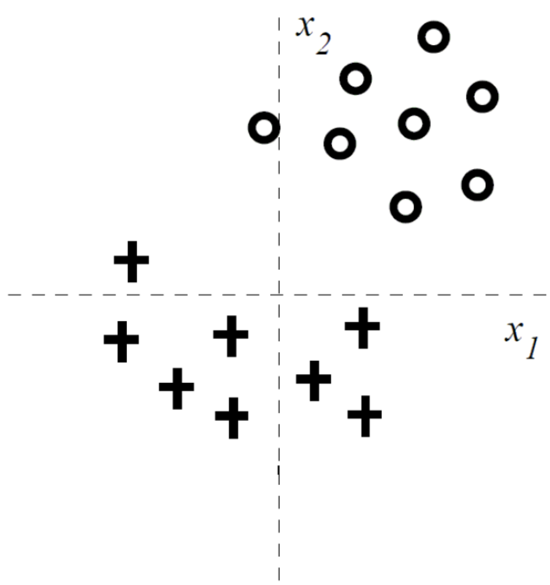

Figure 1: The 2-dimensional labeled training set, where '+' corresponds to class $y = 1$ and 'o' corresponds to class $y = 0$.

Notice that the training data can be separated with zero training error with a linear separator.

Consider training regularized linear logistic regression models where we try to maximize

$$\sum_{i=1}^{N} \log(P(y^i|\boldsymbol{x}^i, w_0, w_1, w_2)) - Cw_j^2$$

for very large C, where N is the number of training samples denoted by $\{(\boldsymbol{x}^i, y^i)\}_{i=1}^N$. The regularization penalties used in penalized conditional loglikelihood estimation are $-Cw_j^2$, where $j = \{0, 1, 2\}$. In other words, only one of the parameters is regularized in each case. Given the training data in Figure 1, how does the training error change with regularization of each parameter w_j? State whether the training error increases or stays the same (zero) for each w_j for very large C. Justify each of your answers.

4) Suppose $X_1, \ldots, X_N$ are i.i.d. samples from the distribution $U(-w, w)$, that is

$$p(x) = \begin{cases} 0 & \text{if } x < -w \\ \frac{1}{2w} & \text{if } |x| \le w \\ 0 & \text{if } x > +w \end{cases}$$

Write down a formula for an MLE estimate of w.

5) Assume we have a binary variable $x \in \{0, 1\}$ with $p(x = 1) \triangleq \eta$. Thus, the variable x has a Bernoulli distribution, i.e., $p(x|\eta) = \eta^x(1 - \eta)^{1-x}$. Our goal is to estimate the value of η given N observations $\{x^{(i)}\}_{i=1}^N$. Assume we have prior information about the parameter η, i.e., we are given $p(\eta)$. Assume η has a Beta distribution with parameters $\alpha, \beta > 0$, i.e.,

$$p(\eta) = \frac{\eta^{\alpha-1}(1-\eta)^{\beta-1}}{B(\alpha, \beta)}, \tag{1}$$

where $B(\alpha, \beta)$ is a normalizing constant. We know that the mode (maximum) of (4) is given by $(\alpha-1)/(\alpha+\beta-2)$. Compute the posterior distribution $p(\eta|x^{(1)}, \ldots, x^{(N)})$ and the MAP estimation of η given the observations.