Machine Learning and Data Mining I (DS 4400)

Midterm II Sample Questions

Instructor: Ehsan Elhamifar

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

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

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

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

3) Consider the problem of separating data $\mathcal{D} = \{(\boldsymbol{x}^1, y^1), \dots, (\boldsymbol{x}^N, y^N)\}$ 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.

4) Consider a binary classification problem in one-dimensional space where the sample contains four data points $S = \{(1, -1), (-1, -1), (2, 1), (-2, 1)\}$ as shown in Fig. 1.

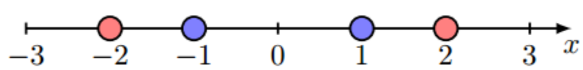

Figure 1: Red points represent instances from class +1 and blue points represent instances from class -1.

A. Define $H_t = [t, \infty)$. Consider a class of linear separateors $\mathcal{H} = \{H_t : t \in \mathbb{R}\}$, i.e., for $\forall H_t \in \mathcal{H}$, $H_t(x) = 1$ if $x \geq t$ otherwise -1. Is there any linear separator $H_t \in \mathcal{H}$ that achieves 0 classification error on this sample? If yes, show one of the linear separators that achieves 0 classification error on this example. If not, briefly explain why there cannot be such linear separator.

B. Now consider a feature map $\phi : \mathbb{R} \to \mathbb{R}^2$ where $\phi(x) = (x, x^2)$. Apply the feature map to all the instances in sample S to generate a transformed sample $S' = \{(\phi(x), y) : (x, y) \in S\}$.Let $\mathcal{H}' = \{ax_1 + bx_2 + c \geq 0 : a^2 + b^2 \neq 0\}$ be a collection of half-spaces in R^2. More specifically, $H_{a,b,c}((x_1, x_2)) = 1$ if $ax_1 + bx_2 + c \geq 0$ otherwise -1. Is there any half-space $H' \in \mathcal{H}'$ that achieves 0 classification error on the transformed sample S'? If yes, give the equation of the max-margin linear separator and compute the corresponding margin.

C. What is the kernel corresponding to the feature map $\phi(\cdot)$ in the last question, i.e., give the kernel function $K(x, z) : \mathbb{R} \times \mathbb{R} \to R$.

5) Consider a two-layer neural network to learn a function $f : X \to Y$, where $X = [X_1, X_2]$ consists of two features. The weights $w_1, \ldots, w_6$ can be arbitrary. There are two possible choices for the function implemented by each unit in this network:

– **S**: sigmoid function, $S(z) = \frac{1}{1+\exp(-z)}$,

– **L**: linear function, $L(z) = cz$,

where in both cases $z = \sum_i w_i X_i$. Assign proper activation functions (**S** or **L**) to each unit in the following graph so that we can generate functions of the form $f(X_1, X_2) = \frac{1}{1+\exp(\beta_1 X_1 + \beta_2 X_2)}$ at the output of the neural network Y. Derive β_1 and β_2 as a function of $w_1, \ldots, w_6$.