

## Homework 1: Review of Probability Statistics and Optimization

**Question 1. (Design of Loss Function)** Let  $X = (X(1), \dots, X(d)) \in \mathbb{R}^d$  and  $Y \in \mathbb{R}$ . In the questions below, make any reasonable assumptions that you need but state your assumptions.

- (a) Prove that  $\mathbb{E}(Y - m(X))^2$  is minimized by choosing  $m(x) = \mathbb{E}(Y | X = x)$ .
- (b) Find the function  $m(x)$  that minimizes  $\mathbb{E}|Y - m(X)|$ . (You can assume that the conditional cdf  $F(y | X = x)$  is continuous and strictly increasing, for every  $x$ .)
- (c) Prove that  $\mathbb{E}(Y - \beta^T X)^2$  is minimized by choosing  $\beta_* = B^{-1}\alpha$  where  $B = \mathbb{E}(XX^T)$  and  $\alpha = (\alpha_1, \dots, \alpha_d)$  and  $\alpha_j = \mathbb{E}(YX(j))$ .

**Question 2. (Central Limit Theorem)** Let  $X_1, \dots, X_n \sim P$ , i.i.d., with  $\mu = \mathbb{E}[X_i]$  and  $\sigma^2 = \text{Var}[X_i]$ . Define

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad s_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

- (i) Prove that  $s_n^2 \xrightarrow{P} \sigma^2$ .
- (ii) Prove that  $\sqrt{n}(\bar{X}_n - \mu)/s_n \xrightarrow{d} N(0, 1)$ .  
*(hint: using Slutsky's Theorem [https://en.wikipedia.org/wiki/Slutsky%27s\\_theorem](https://en.wikipedia.org/wiki/Slutsky%27s_theorem))*

**Question 3. (Curse of Dimensionality: Asymptotic scaling of nearest neighbor distances)**

- (a) Let  $x_0, x_1, \dots, x_n$  be i.i.d. from a distribution  $P$  supported on  $[-R, R]^d$ . Let  $i(x_0)$  be the index of the closest point (in  $\ell_2$  distance) among  $x_{1:n} = \{x_1, \dots, x_n\}$  to  $x_0$ . Prove that for any  $\delta > 0$ ,

$$\mathbb{P}(\|x_{i(x_0)} - x_0\|_2 > \delta) = \int (1 - P(B_d(x, \delta)))^n dP(x),$$

where  $B_d(x, \delta)$  denotes the  $\ell_2$  ball of radius  $\delta$  centered at  $x$ . To be clear, the probability on the left-hand side above is over  $x_0$  and  $x_{1:n}$ .

- (b) Prove that for any  $\delta$ , there exists a rectangular partition  $U_1, \dots, U_{N(\delta)}$  of  $[-R, R]^d$  with diameter at most  $\delta$ , and

$$N(\delta) \leq \frac{c}{\delta^d},$$

where  $c > 0$  is a constant depending only on  $R$  and  $d$ .

- (c) Using parts (a) and (b), prove that

$$\mathbb{P}(\|x_{i(x_0)} - x_0\|_2 > \delta) \leq \frac{c}{en\delta^d}.$$

*Hint: first show that*

$$\mathbb{P}(\|x_{i(x_0)} - x_0\|_2 > \delta) \leq \sum_{j=1}^{N(\delta)} \int_{U_j} (1 - P(U_j))^n dP(x) = \sum_{j=1}^{N(\delta)} P(U_j) (1 - P(U_j))^n.$$

*Then show that each summand above is bounded by  $1/(en)$ .*

- (d) Argue that the last part translates to

$$\|x_{i(x_0)} - x_0\|_2 \lesssim \left(\frac{1}{n}\right)^{1/d} \quad \text{in probability.}$$

**Question 4. (Duality of Support Vector Machine)** Consider a training dataset  $\mathcal{D} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots\}$ . We distinguish between two types of supervised learning problems depending on the targets  $y^{(i)}$ . Let's consider the Binary Classification problem where the target variable  $y$  is discrete and takes on one of  $K = 2$  possible values. (we assume  $\mathcal{Y} = \{-1, +1\}$ .) We will also work with linear models of the form:

$$f_{\theta}(x) = \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_d \cdot x_d$$

where  $x \in \mathbb{R}^d$  is a vector of features and  $y \in \{-1, 1\}$  is the target. The  $\theta_j$  are the parameters of the model. We can represent the model in a vectorized form  $f_{\theta}(x) = \theta^{\top} x + \theta_0$ . Next we define the *geometric margin*  $\gamma^{(i)}$  with respect to a training example  $(x^{(i)}, y^{(i)})$  as

$$\gamma^{(i)} = y^{(i)} \left( \frac{\theta^{\top} x^{(i)} + \theta_0}{\|\theta\|} \right).$$

(a) Show that this corresponds to the distance from  $x^{(i)}$  to the hyperplane.

(b) We saw that maximizing the margin of a linear model amounts to solving the following optimization problem.

$$\min_{\theta, \theta_0} \frac{1}{2} \|\theta\|^2$$

subject to

$$y^{(i)} \left( (x^{(i)})^{\top} \theta + \theta_0 \right) \geq 1 \text{ for all } i$$

write down the Lagrangian of the max-margin optimization problem.

**Hint:** convex duality theory: <https://web.stanford.edu/class/ee364a/lectures/duality.pdf>

(c) An interesting question arises when we need to decide which optimization problem to solve: the dual or the primal.

NORTHWESTERN UNIVERSITY