Machine Learning and Data Mining I (DS 4400)

Instructor: Ehsan Elhamifar

Midterm 1 Sample Questions

First, Last Name:

Husky Email:

1) Regression (40 pts). Assume we have a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ denote the input and the output, respectively, corresponding to the i-th observation. The goal is to learn a function $h_{\boldsymbol{\theta}}(\cdot)$ so that $h_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ is as close as possible to y^i.

In linear regression, we choose $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \phi(\boldsymbol{x})$, where $\phi : \mathbb{R}^d \to \mathbb{R}^p$ is the basis function expansion. We obtain the optimal parameters $\hat{\boldsymbol{\theta}}$ by minimizing the total loss function

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ell(h_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i), \tag{1}$$

where $\ell(\cdot, \cdot)$ is a cost function, e.g., $\ell(h_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i) = (h_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)^2$.

1. (10 pts) Provide a robust cost function $\ell(h_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)$ which can discount the the effect of outliers. Explain why it provides robustness to outliers.

2. (10 pts) What is overfitting (explain)? Explain which of the following cases is more likely to suffer from overfitting.

   (a) $p = 30, N = 25$

   (b) $p = 20, N = 30$

3. (20 pts) Consider the regularized least squares problem

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left(\boldsymbol{\theta}^\top \phi(\boldsymbol{x}_i) - y_i\right)^2 + \lambda \|\boldsymbol{\theta} - \boldsymbol{a}\|_2^2, \tag{2}$$

   where $\boldsymbol{a}$ is a known and given vector of the same dimension as $\boldsymbol{\theta}$. Derive the closed-form solution for (2). Provide all steps of the derivation.

2. Maximum Likelihood Estimation (20 pts). Assume $X_1, X_2, \ldots, X_N$ are i.i.d. random variables each taking a real value, where

$$p_\delta(X_i = x_i) = e^{-(\delta^2 + \delta x_i)}.$$

Here, δ is the parameter of the distribution. Assume, we observe $X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N$.

1. (10 pts) Write down the likelihood function $L(\delta)$.

2. (10 pts) Derive the maximum likelihood estimation of δ for the given observations. Provide all steps of derivations.

3. Logistic Regression (20 pts). In the logistic regression for binary classification ($y \in \{0, 1\}$), we defined $p(y = 1|x) = \sigma(\boldsymbol{w}^\top \boldsymbol{x})$, where the sigmoid function is defined as

$$\sigma(z) \triangleq \frac{1}{1 + e^{-z}}.$$

Assume we have trained the logistic regression model using a given dataset and have learned $\boldsymbol{w}$. Let $\boldsymbol{x}_n$ be a test sample.

1. (10 pts) Assume $\boldsymbol{w}^\top \boldsymbol{x}_n < 0.2$. To which class $\boldsymbol{x}_n$ belongs? Provide all details of your derivations.

2. (10 pts) Assume $\frac{1}{1+e^{\boldsymbol{w}^\top \boldsymbol{x}_n}} = 0.8$. To which class $\boldsymbol{x}_n$ belongs and with what probability? Provide all details of your derivations.

4. Probability and Linear Algebra (20 pts).

1. (10 pts) Assume X_1, X_2, X_3 are three random variables, where X_1 and X_3 are independent. Show that the following holds:

$$P(X_1, X_2, X_3) = P(X_1)P(X_2|X_1, X_3)P(X_3).$$

2. (10 pts) Assume $f(x) = g(x^2)$, where $g(\cdot)$ is a differentiable and monotone increasing function, i.e., $g'(z) \geq 0$ for every z. Investigate if $f(x)$ is a convex function. If it is, prove it, if it os not, show by a counterexample (e.g., choose a g for which f is non-convex).