Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2021

Writing Assignment 3

Issued: Friday 12th November, 2021 **Due:** Friday 26th November, 2021

3.1. (4 points) Important inequalities in Learning Theory.

(a) (2 points) (Markov's Inequality) Let X be a non-negative random variable, then for every positive constant a, please show that

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

(b) (2 points) (Chebyshev's inequality)For random variable X, if its expected value $\mathbb{E}(X)$ and variance $Var(X)$ are both finite, for every positive constant a, please show that

$$P(|X - \mathbb{E}(X)| \geq a) \leq \frac{Var(X)}{a^2}.$$

3.2. (4 points) (Regularization) This problem explores a statistical motivation to l_1-norm regularized linear regression, a.k.a LASSO.

Given m samples $(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(m)}, y^{(m)}), \boldsymbol{x}^{(i)} \in \mathbb{R}^n, y \in \mathbb{R}, i = 1, \cdots, m$, we need to determine the parameters $\boldsymbol{\theta} \in \mathbb{R}^n$ for the linear model:

$$y^{(i)} = \boldsymbol{\theta}^\top \boldsymbol{x}^{(i)} + \epsilon^{(i)},$$

where $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. Gaussian random variables. Unlike Maximum Likelihood Estimation, we treat $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_n]^{\mathrm{T}}$ as a random variable with prior laplace distribution:

$$P(\theta_i) = \frac{1}{2}\exp(-|\theta_i|), i = 1, \ldots, n.$$

The **Maximum A Posteriori (MAP)** estimation of parameter $\boldsymbol{\theta}$ is defined as follows:

$$\boldsymbol{\theta}_{MAP} \triangleq \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} P(y^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}) \cdot \prod_{j=1}^{n} P(\theta_j).$$

Please show that LASSO is equivalent to the MAP estimation of this model.

3.3. (4+4 points) (Back Propagation) Consider the back propagation on the hidden-layer in a neural network. Given a batch of input feature $X = [x^{(1)}, x^{(2)}, \cdots, x^{(M)}]^T$ (Shape: $M \times D_0$), a set of weight $\{W \in \mathbb{R}^{D_0 \times D_1}, b \in \mathbb{R}^{D_1 \times 1}\}$, and element-wise Sigmoid activation function $\sigma(\cdot)$. The forward propagation on this hidden-layer is given by:

$$F_1 = XW + \mathbb{1}_M b^T, \quad F_2 = \sigma(F_1)$$

where $\mathbb{1}_M$ is a vector composed of 1 in length M. $\sigma(\cdot)$ is element-wise Sigmoid function:

$$[\sigma(X)]_{ij} = \frac{1}{1 + \exp(-X_{ij})}$$

Then in the back propagation stage, suppose we already know the gradients for some scalar loss function l with respect to F_2 as $\nabla_{F_2} l$.

(a) (4 points) Proceed the back propagation and show that $\nabla_{F_1} l = (\nabla_{F_2} l) \odot F_2 \odot (1 - F_2)$. where $(A \odot B)_{ij} = A_{ij} B_{ij}$ is element-wise production.

(b) (4 points) (Bonus) Prove that

$$\begin{aligned}
\nabla_X l &= (\nabla_{F_2} l) \odot F_2 \odot (1 - F_2) \cdot W^{\mathrm{T}}, \\
\nabla_W l &= X^{\mathrm{T}} \cdot (\nabla_{F_2} l) \odot F_2 \odot (1 - F_2), \\
\nabla_b l &= \big((\nabla_{F_2} l) \odot F_2 \odot (1 - F_2)\big)^{\mathrm{T}} \cdot \mathbb{1}_M\,.
\end{aligned}$$