CS 6140: Machine Learning — Fall 2021— Paul Hand

HW 5 – REVISED

Due: Wednesday October 20, 2021 at 2:30 PM Eastern time via Gradescope.

Names: [Put Your Name(s) Here]

You can submit this homework either by yourself or in a group of 2. You may consult any and all resources. Make sure to justify your answers. If you are working alone, you may either write your responses in LaTeX or you may write them by hand and take a photograph of them. If you are working in a group of 2, you must type your responses in LaTeX. You are encouraged to use Overleaf. Create a new project and replace the tex code with the tex file of this document, which you can find on the course website. To share the document with your partner, click Share > Turn on link sharing, and send the link to your partner. When you upload your solutions to Gradescope, make sure to take each problem with the correct page or image.

Question 1. *Log loss with multiclass regression.*

Consider training data $\{(x_i, y_i)\}_{i=1...n}$ where $x_i \in \mathcal{X} = \mathbb{R}^d$ and $y_i \in \mathcal{Y} = \{1 \dots K\}$. Suppose that the data are given by a distribution where $y \mid x \sim f(x;\theta)$, where $f(x;\theta)$ is a probability distribution over $\mathcal{Y}$. That is, $P(y = k \mid x;\theta) = f(x;\theta)_k$. Show that the maximum likelihood estimate of θ is given by solving

$$\min_\theta - \sum_{i=1}^n \sum_{k=1}^K 1_{y_i=k} \cdot \log f(x_i;\theta)_k$$

Here, $1_A = \begin{cases} 1 & \text{if } A, \\ 0 & \text{otherwise.} \end{cases}$ is an *indicator function*.

Response:

Question 2. *Linear Regression with non-Gaussian noise*

Consider training data $\{(x_i, y_i)\}_{i=1...n}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Suppose $y_i = x_i^t\theta + \varepsilon_i$, where ε_i follows a Laplace distribution with mean zero, and $\theta \in \mathbb{R}^d$. Show that the maximum likelihood estimate of θ is given by solving

$$\min_\theta \sum_{i=1}^n |y_i - x_i^t\theta|$$

Response:

Question 3. *Risk and linear regression*

Consider training data $\{(x_i, y_i)\}_{i=1...n}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Suppose $x \sim \mathcal{N}(0, I_d)$, where I_d is the $d \times d$ identity matrix. Suppose $y_i = x_i^t\theta^* + \varepsilon_i$, where $\theta^* \in \mathbb{R}^d$ and ε_i is a zero-mean random variable with variance σ^2, independent of x_i.

(a) Consider a predictor given by $f_\theta : \mathbb{R}^d \to \mathbb{R}$, with $f_\theta(x) = x^t\theta$. Show that the risk of f_θ with respect to the square loss is

$$\mathcal{R}(f_\theta) = \|\theta^* - \theta\|^2 + \sigma^2.$$

Response:

(b) Show that the solution to least squares linear regression, $\hat{\theta}$ is unbiased in the sense that $\mathbb{E}(\hat{\theta}) = \theta^*$.

Response:

Additional resources/hints that may help:

Problem 1:

- Log loss function math explained
- Hint: use indicators as exponents when setting up the likelihood
- Hint: you do not need to deal with one-hot encodings to do this problem

Problem 3:

- Multivariate Gaussian Distribution, Covariance Matrix
- Bias of an estimator
- Law of Iterated Expectation