**EC4213/ET5402/CT5303: Machine learning and deep learning** **Fall 2018**

# Homework 01

(due October 8, 2018)

*Instructor: Jonghyun Choi* *Student name (ID#):*

Please specify your name and your student ID in the top heading. Hand in as a fully answered version to github classroom.

## Problem 1 : Maximum Likelihood Estimation (30 pt)

Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a statistical model given observations, by finding the parameters that maximize the likelihood of the observations. Concretely, given observations $y_1, y_2, ..., y_n$ distributed according to $p_\theta(y_1, y_2, ..., y_n)$ (here $p_\theta$ can be a probability mass function for discrete observations or a density for continuous observations), the likelihood function is defined as $L(\theta) = p_\theta(y_1, y_2, ..., y_n)$ and the MLE is

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, L(\theta)$$

We often make the assumption that the observations are independent and identically distributed or iid, in which case $p_\theta(y_1, y_2, ..., y_n) = f_\theta(y_1) f_\theta(y_2) \cdot \cdot \cdot \cdot \cdot f_\theta(y_n)$.

(a) The instructor recommends maximizing the log-likelihood $\ell(\theta) = \log L(\theta)$ instead of maximizing $L(\theta)$. Why does this yield the same solution $\hat{\theta}_{MLE}$?

(b) Why is it easier to solve the optimization problem for $\ell(\theta)$ in the iid case?

(c) The Poisson distribution is $f_\lambda(y) = \frac{\lambda^y e^{-\lambda}}{y!}$. Let $Y_1, Y_2, ..., Y_n$ be a set of independent and identically distributed random variables with Poisson distribution with parameter $\lambda$. Find the joint distribution of $Y_1, Y_2, ..., Y_n$.

(d) Find the maximum likelihood estimator of $\lambda$ as a function of observations $y_1, y_2, ..., y_n$.

## Problem 2 : The accuracy of learning decision boundaries (40 pt)

This problem exercises your basic probability in the context of understanding why lots of training data helps to improve the accuracy of learning things.

For each $\theta \in (\frac{1}{3}, \frac{2}{3})$, define $f_\theta : [0, 1] \to \{0, 1\}$, such that

$$f_\theta(x) = \begin{cases} 1 & \text{if } x > \theta \\ 0 & \text{otherwise.} \end{cases}$$

*(Hint : draw the plot of $f_\theta(x)$ w.r.t. $x$ may help you to solve the problems.)*

We draw samples $X_1, X_2, ..., X_n$ uniformly at random and i.i.d. from the interval $[0, 1]$. Our goal is to learn an estimate for $\theta$ from $n$ random samples $(X_1, f_\theta(X_1)), (X_2, f_\theta(X_2)), ..., (X_n, f_\theta(X_n))$.

— Let $T_{min} = \max(\{\frac{1}{3} \cup \{X_i | f_\theta(X_i) = 0\})$. We know that the true $\theta$ must be larger than $T_{min}$.

— Let $T_{max} = \min(\{\frac{2}{3} \cup \{X_i | f_\theta(X_i) = 1.\})$. We know that the true $\theta$ must be smaller than $T_{max}$.

The gap between $T_{min}$ and $T_{max}$ represents the uncertainty we will have about the true $\theta$ given the training data that we have received.

(a) What is the probability that $T_{max}$ - $\theta > \alpha$ as a function of $\alpha$ ?

(b) What is the probability that $\theta$ - $T_{min} > \alpha$ as a function of $\alpha$ ?

(c) Suppose that you would like the estimator $\hat{\theta} = (T_{max} + T_{min})/2$ for $\theta$ that is $\alpha$-close (defined as $|\hat{\theta} - \theta| < \alpha$, where $\hat{\theta}$ is the estimation and $\theta$ is the true value) with probability at least $1 - \delta$. Both $\alpha$ and $\delta$ are some small positive numbers. Please bound or estimate how big of an $n$ do you need ? (**Note** : You do not need to find the optimal lowest sample complexity $n$, an approximation using results of question (a) is fine.)

## Problem 3 : Geometry of Ridge Regression (20 pt)

You recently learned ridge regression and how it differs from ordinary least squares. In this question we will explore how ridge regression is related to solving a constrained least squares problem in terms of their parameters and solutions.

Recall that ridge regression is given by the unconstrained optimization problem

$$\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_2^2. \tag{1}$$

(a) Derive that the solution to ridge regression (1) is given by $\hat{\mathbf{w}}_\mathbf{r} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$.

(b) In the solution to the above problem, What happens when $\lambda \to \infty$ ? (Hint : It is for this reason that sometimes regularization is referred to as "shrinkage".)