EE488C Special Topics in EE <Deep Learning and AlphaGo> October 2, 2016
EE, KAIST, Fall 2016 Due: 1:00pm, October 10

Problem Set #2

Warning: Homeworks will not be graded if submitted after the deadline. For all problems, show detailed reasoning.

0. (Reading assignment) DL Book Chapters 6 $\sim$ 7

1. (Bias-variance tradeoff) Consider a problem of estimating θ from a set of samples $\{x^{(1)}, \ldots, x^{(m)}\}$ generated i.i.d. according to Bernoulii(θ), where $0 \leq \theta \leq 1$. As shown in (5.28) in page 125 of the textbook, the sample-mean estimator $\hat{\theta}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$ given in (5.22) is unbiased, i.e., bias$(\hat{\theta}_m) = \mathbb{E}[\hat{\theta}_m] - \theta = 0$. The variance of the sample-mean estimator is $\frac{1}{m}\theta(1-\theta)$ as shown in (5.52) in page 129.

(a) Let's consider a trivial estimator $\tilde{\theta}_m = 1$, i.e., it always says 1 regardless of the samples it observes. Calculate the bias and variance of the trivial estimator.

(b) Assume $m = 1$. Can you think of an estimator that minimizes the worst-case MSE over all θ? Namely, find an estimator $\bar{\theta}_1$ based on $x^{(1)}$ that minimizes the following.

$$\max_{0\leq\theta\leq 1} \mathbb{E}[(\bar{\theta}_1 - \theta)^2]$$

What is the minimum worse-case MSE that the estimator achieves?

(c) Assume $m = 1$ and plot the mean squared error (MSE)[1], the squared bias, and the variance for all three estimators, i.e., the sample-mean estimator, the trivial estimator in (a), and the estimator you found in (b) for all values of $0 \leq \theta \leq 1$. For which values of θ, do you observe a bias-variance tradeoff similar to Fig. 5.6, i.e., the trivial estimator is in the underfitting zone, the sample-mean estimator is in the overfitting zone, and the optimal estimator you found in (b) is in between the two?

(d) Assume $m = 1$. Can you think of an estimator that minimizes the average MSE over all θ? Namely, find an estimator $\bar{\theta}_1$ based on $x^{(1)}$ that minimizes the following.

$$\int_0^1 \mathbb{E}[(\bar{\theta}_1 - \theta)^2] d\theta$$

How is this different from the estimator in (b)?

2. (Linear regression for Problem 1) Let's consider Problem 1 as a linear regression problem, i.e., we want to build a model to output $\hat{y} = \mathbf{w}^T\mathbf{x} = \sum_{i=1}^{n} w_i x_i$ so that it gives a good estimate of θ, where $\{x_1, \ldots, x_n\}$ are generated i.i.d. Bernoulli(θ). Assume $0 < \theta < 1$.

[1]Note that the MSE is equal to $\text{Bias}(\hat{\theta}_m)^2 + \text{Var}(\hat{\theta}_m)$ as shown in (5.54) because

$$\begin{aligned}
\text{MSE} &= \mathbb{E}[(\hat{\theta}_m - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta}_m - \mathbb{E}(\hat{\theta}_m) + \mathbb{E}(\hat{\theta}_m) - \theta)^2] \\
&= \text{Var}(\hat{\theta}_m) + 2\mathbb{E}[\hat{\theta}_m - \mathbb{E}(\hat{\theta}_m)]\mathbb{E}[\mathbb{E}(\hat{\theta}_m) - \theta] + (\mathbb{E}(\hat{\theta}_m) - \theta)^2 \\
&= \text{Var}(\hat{\theta}_m) + 2\{\mathbb{E}(\hat{\theta}_m) - \mathbb{E}(\hat{\theta}_m)\} \cdot \{\mathbb{E}(\hat{\theta}_m) - \theta\} + \text{Bias}(\hat{\theta}_m)^2 \\
&= \text{Var}(\hat{\theta}_m) + \text{Bias}(\hat{\theta}_m)^2.
\end{aligned}$$

(a) Let's generate a training set consisting of m examples with a fixed θ. Then, the training set is given by $(\mathbf{X}, \mathbf{y})$, where $\mathbf{X}$ is an $m \times n$ matrix whose elements are all i.i.d. Bernoulli(θ) random variables and $\mathbf{y}$ is an $m \times 1$ vector whose entries are all equal to θ. Find the optimum $\mathbf{w}^*$ that minimizes the mean square error MSE $= \frac{1}{m} \|\hat{\mathbf{y}} - \mathbf{y}\|^2$. Will it be equivalent to the sample-mean estimator in Problem 1 as $m, n \to \infty$?[2] You don't need to perform rigorous analysis for convergence of random variables. For example, you may use the approximations $\frac{1}{m} \sum_{i=1}^{m} X_{i,j}^2 \sim \theta$, $\forall j$ and $\frac{1}{m} \sum_{i=1}^{m} X_{i,j} X_{i,k} \sim \theta^2$, $\forall j \neq k$, which become accurate as $m \to \infty$.

(b) Now, let's assume $n = 1$ and generate m examples each time choosing θ uniformly between 0 and 1.[3] Furthermore, let's introduce a bias term in the model, i.e., the model is now $\hat{y} = wx + b$. Find the optimum w and b that minimizes the MSE $\frac{1}{m} \|\hat{\mathbf{y}} - \mathbf{y}\|^2$ as $m \to \infty$. Note that you can still use (5.12) to solve this problem by assuming $\hat{y} = w_1 x + w_2 x_2$, where x_2 is fixed to 1. Then, $\mathbf{X}$ is an $m \times 2$ vector, where its i-th entry in the first column is Bernoulli(θ_i) and its second column is all 1 and $\mathbf{y}$ is an $m \times 1$ vector whose i-th element is θ_i, where θ_i's are i.i.d. Uniform($[0, 1]$). How does this compare with the estimator you found in Problem 1 (d)?

3. (Unsupervised learning) Design a machine learning algorithm that can automatically classify {A,B,C,D,E,H,I,K,M,N,O,S,T,U,V,W,X,Y,Z}, into the following four categories.

- A,M,T,U,V,W,Y
- B,C,D,E,K
- N,S,Z
- H,I,O,X

4. (Deep learning for XOR) In page 26 of Lecture notes #7, we saw that the following point becomes a global minimum if $a = 1$, but it is locally convex as a changes in the neighborhood of -1.[4] Is the following point with $a = -1$ a local minimum or a saddle point?

$$\mathbf{W} = \begin{pmatrix} a & -a \\ -a & 1 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} 0 \\ a - 1 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad b = 0$$

[2] Note that since the correct answer θ is already given in the training set for this problem, an optimum estimator is the one that simply outputs θ. However, the goal of this problem is to see whether a linear regression can learn to solve the problem.

[3] In Problem 1, the number of samples was m, which corresponds to n in Problem 2. m in Problem 2 is the number of examples, where each example n dimensional.

[4] There was a typo in page 26, i.e., $\mathbf{c}$ should be $(0, a-1)^T$ not $(0, 0)^T$.