Name: Honda Li

Short Answer and "True or False" Conceptual Questions

A.0

(A.0.a)

The bias and variance is similar to the concepts of precision and accuracy in Experimental Physics. Under the context of machine learning, Bias refers to part of the learning errors caused by a model being too simple, in a way that it just cant represent the joint probability density function with its simplicity.

The variance refers to the variance of the random variable $\hat{f}(x)$, which depends on the samples we observed.

Bias-variance trade off relates 2 types of learning errors (Bias, Variance) with the model complexity.

(A.0.b)

Usually, higher the model complexity, higher the variance, lower the model complexity, higher the bias.

(A.0.c)

False. The bias decreases. Because the bias is: $\mathbb{E}\left[\left(f(x) - \mathbb{E}\left[\hat{f}(x)\right]\right)^2\right]$. The number of sample seems to be irrelevant to the amount of bias we have for the model.

(A.0.d)

True when we fix the model complexity. This is absolutely true when we consider a linear regression. If we take infinitely many data, the best line-fit will converge.

This is even more true when we just consider taking the average as a way of making prediction (Which is not a bad way to predict the output given $X = x$), then, this is literally the Central limit theorem, as we have more and more samples, the variance of the sample average gets smaller.

(A.0.e)

(A.0.f)

We should use the test set to tune the hyper-parameters for the models.

(A.0.g)

False. It's an underestimate.

Maximum Likelihood Estimator(MLE)

(A.1)

(A.1.a)

Objective: Find the expression for the maximum=likelihood estimate for the parameter λ for the poisson distribution, interns of the goal count. Assume idd rvs.

Here we will assume that observations obtained takes the form $x_1, x_2, \cdots x_N$, and then we derive the best estimator for λ in this much general context.

I will shut up and just show you the math:

$$\prod_{n=1}^{N} \text{Poi}(x_n|\lambda) \tag{A.1.a.1}$$

$$\sum_{n=1}^{N} \log\left(\text{Poi}(x_n|\lambda)\right)$$

$$\sum_{n=1}^{N} \left(-\lambda + x_n \ln(\lambda) + \log(x_n!)\right)$$

Notice that, only some of the terms are relevant to the parameter λ, therefore, the optimization problem we are solving is:

$$\lambda^+ = \underset{\lambda}{\text{argmax}} \left\{ -N\lambda + \ln(\lambda) \sum_{i=1}^{N} (x_n) \right\} \tag{A.1.a.2}$$

To solve it, we just take the derivative, set it to zero and then solve for λ, because this function is a function that has a single local maximum.

$$\partial_\lambda \left[-N\lambda + \ln(\lambda) \sum_{i=1}^{N} (x_n) \right] = 0 \tag{A.1.a.3}$$

$$-N + \frac{\sum_{n=1}^{N} x_n}{\lambda^+} = 0$$

$$\implies \lambda^+ = \frac{\sum_{n=1}^{N} x_n}{N}$$

Therefor, for this particular problem, the best estimator will be the average of all the observation, which is just:

$$\frac{2+4+6+1}{5} = 2.6$$

And that is the answer for the question.

(A.1.b)

The derivation of the best estimator in the general context is shown in A.1.a.
The numerical value for six observations is:

$$\frac{2+4+6+1+3}{7} = 2.6666666666\cdots$$

(A.1.c)

The numerical results for 5 observations has been shown in A.1.a and A.1.b respectively.

(A.2)

Objective: Find the MLE for the uniform distribution on $[o, \theta]$, where θ is the value we want to estimate.

Suppose that an observations has been made: $x_1, x_2, \cdots x_N$, and we assume that they are idd, and we want to find the likelihood of such an observation is generated using the Uniform distribution. And this will be given by the expression:

$$\prod_{n=1}^{N} \underbrace{\mathbb{P}\left(X = x_n\right)}_{\frac{1}{\theta} \mathbf{1}\{0 \leq x_n \leq \theta\}} \tag{A.2.1}$$

Observe that, if any of the observation is beyond the range $[0, \theta]$, we will have zero likelihood, so let's assume that $\theta \leq \max_{1 \leq i \leq N}(x_i)$, then we will have this expression for the likelihood:

$$\prod_{n=1}^{N} \frac{1}{\theta} = \frac{1}{\theta^N} \tag{A.2.2}$$

And taking the log on that we have:

$$\log\left(\frac{1}{\theta^N}\right) = -N \log(\theta) \tag{A.2.3}$$

Observe that that function is monotonically decreasing as the value of θ get larger and larger, therefore, to maximize the likelihood, we need the value of θ to be as small as possible, and the smallest possible such θ that is not giving us zero likelihood is: $\max_{1 \leq i \leq N}(x_i)$, therefore, best estimate is given by:

$$\theta^+ = \max_{1 \leq i \leq N}(x_i)$$

Over-fitting

A.3

A.3.a

A.3.b

A.3.c

Polynomial Regression

Ridge Regression on MNIST