DS 100: Principles and Techniques of Data Science **Date: March 16, 2018**

Discussion #7

Name:

Bias-Variance Tradeoff

1. Let X be a random variable with mean $\mu = \mathbb{E}[X]$. Using the definition $\text{Var}(X) = \mathbb{E}[(X-\mu)^2]$, show that for any constant c,

$$\mathbb{E}[(X-c)^2] = (\mu - c)^2 + \text{Var}(X).$$

Solution: One way to show this is to write $X - c = X - \mu + \mu - c$. Squaring both sides,

$$\mathbb{E}[(X-c)^2] = \mathbb{E}[(X-\mu)^2 + (\mu - c)^2 + 2(X-\mu)(\mu - c)]$$

Now using linearity of expectation and pulling out the constants,

$$\begin{aligned}\mathbb{E}[(X-c)^2] &= \mathbb{E}[(X-\mu)^2] + (\mu - c)^2 + 2\underbrace{\mathbb{E}[X-\mu]}_{=0}(\mu - c) \\ &= \text{Var}(X) + (\mu - c)^2.\end{aligned}$$

2. In the context of question 1, conclude that

- $\text{Var}(X) \le \mathbb{E}[(X-c)^2]$ for any c
- $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

Solution: The first bullet follows from using $(\mu - c)^2 \ge 0$, and the second bullet follows from plugging in $c = 0$.

3. Suppose we make **independent** observations $X_1, \dots, X_n$ with a common density $f(x)$, and we construct a KDE to estimate the density:

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i),$$

where $K_h(y) = K(y/h)/h$.

(a) Write the bias-variance decomposition for the $\mathbb{L}_2$-error $\mathbb{E}[(\widehat{f}(x) - f(x))^2]$ at a point x.

> **Solution:** Note $\widehat{f}(x)$ is random and $f(x)$ is fixed.
>
> $$\mathbb{E}[(\widehat{f}(x) - f(x))^2] = \left(\mathbb{E}[\widehat{f}(x)] - f(x)\right)^2 + \text{var}[\widehat{f}(x)].$$

(b) What happens to each term as the number of samples n increases?

> **Solution:** Note that the expected value
>
> $$\mathbb{E}\left[\widehat{f}(x)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[K_h(x - X_i)\right] = \mathbb{E}[K_h(x - X_1)],$$
>
> does not depend on n, so the bias does not depend on the number of samples. As we showed in lecture,
>
> $$\text{var}[\widehat{f}(x)] = \frac{\text{var}[K_h(x - X_1)]}{n},$$
>
> so the variance decreases with the number of samples.

(c) What happens to each term as the bandwidth h approaches 0 or ∞?

> **Solution:** $\mathbb{E}\left[\widehat{f}(x)\right] = \mathbb{E}[K_h(x - X_1)]$. If f is a probability mass function, we can write out this expectation as a sum over the possible values $\mathcal{X}$ that X_1 can take on:
>
> $$\mathbb{E}[K_h(x - X_1)] = \sum_{t\in\mathcal{X}} f(t)K_h(x - t).$$
>
> When $h \to 0$, the term $K_h(x - t)$ is really small unless x is very close to t, so $\lim_{h\to 0}\mathbb{E}\left[\widehat{f}(x)\right] = \lim_{h\to 0}\mathbb{E}[K_h(x - X_1)] = f(x)$. When h approaches ∞, $K_h(x-t)$ is tiny everywhere (the kernel looks flat), so $\sum_{t\in\mathcal{X}} f(t)K_h(x - t) \approx 0$.
>
> For the variance, when $h \to \infty$ the estimate $\widehat{f}(x)$ does not depend on the data at all, so it should be very low variance indeed! When $h \to 0$, the kernel $K_h(x - X_1)$ is puts its mass around X_1. If the value of X_1 is noisy (i.e. changes a lot across draws), where we put any mass at all will vary a lot too.

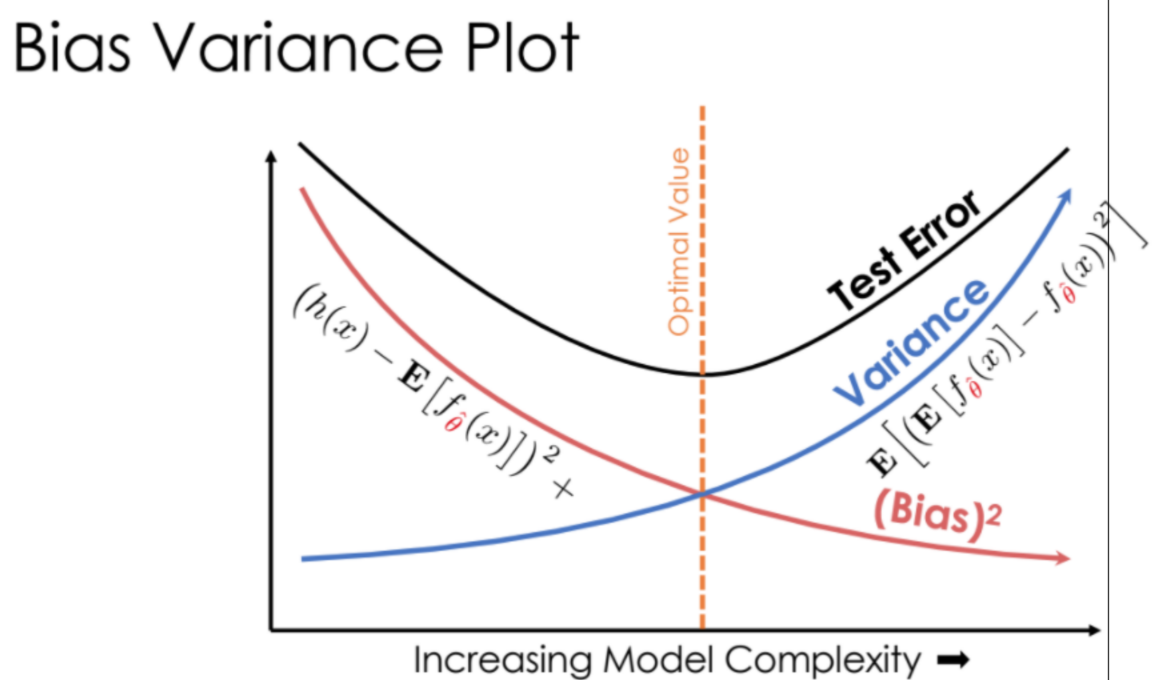
4. Recall that we can break down squared error into Noise, Bias and Variance:

$$\mathbb{E}\left((y - f(x))^2\right) = \mathbb{E}\left[(y - h(x))^2\right] + (h(x) - \mathbb{E}(f(x)))^2 + \mathbb{E}\left[(\mathbb{E}(f(x)) - f(x))^2\right]$$

where $y = h(x) + \epsilon$, $\mathbb{E}(\epsilon) = 0$, Var$(\epsilon) = \sigma^2$

As we increase model complexity, how are these terms affected? Draw a graph showing how variance, bias and test error change as model complexity increases.

Solution: As we increase model complexity, the bias term decreases, while the variance term increases.



Regularization

5. In a petri dish, yeast populations grow exponentially over time. In order to estimate the growth rate of a certain yeast, you place yeast cells in each of n petri dishes and observe the population y_i at time x_i and collect a dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Because yeast populations are known to grow exponentially, you propose the following model:

$$\log(y_i) = \beta x_i \tag{1}$$

where β is the growth rate parameter (which you are trying to estimate). We will derive the L_2 regularized estimator least squares estimate.

(a) Write the *regularized least squares loss function* for β under this model. Use λ as the regularization parameter.

Solution:

$$L(\beta) = \frac{1}{n}\sum_{i=1}^{n}(\log(y_i) - \beta x_i)^2 + \lambda\beta^2 \tag{2}$$

(b) Solve for the optimal $\widehat{\beta}$ as a function of the data and λ.

Solution: Taking the derivative of the regularized loss function function:

$$\begin{aligned}
\frac{\partial}{\partial\beta}L(\beta) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta}(\log(y_i) - \beta x_i)^2 + \frac{\partial}{\partial\beta}\lambda\beta^2 && (3)\\
&= -\frac{2}{n}\sum_{i=1}^{n}(\log(y_i) - \beta x_i)x_i + 2\lambda\beta && (4)\\
&= -\frac{2}{n}\sum_{i=1}^{n}\log(y_i)x_i + \frac{2}{n}\sum_{i=1}^{n}\beta x_i^2 + 2\lambda\beta && (5)\\
&= -\frac{2}{n}\sum_{i=1}^{n}\log(y_i)x_i + \frac{2}{n}\beta\left(\sum_{i=1}^{n}x_i^2 + \lambda n\right) && (6)\\
& && (7)
\end{aligned}$$

Setting the derivative equal to zero and solving for β:

$$0 = -\frac{2}{n}\sum_{i=1}^{n}\log(y_i)x_i + \frac{2\beta}{n}\left(\lambda n + \sum_{i=1}^{n} x_i^2\right) \tag{8}$$

$$\beta\left(\lambda n + \sum_{i=1}^{n} x_i^2\right) = \sum_{i=1}^{n}\log(y_i)x_i \tag{9}$$

$$\beta = \left(\lambda n + \sum_{i=1}^{n} x_i^2\right)^{-1}\sum_{i=1}^{n}\log(y_i)x_i \tag{10}$$

$$\tag{11}$$