DS 100/200: Principles and Techniques of Data Science **Date: October 30, 2019**

Discussion # 10 Solutions

Name:

Bias-Variance Trade-off

1. Assume that we have a function $h(x)$ and some noise generation process that produces ϵ such that $\mathbb{E}[\epsilon] = 0$ and $\text{var}(\epsilon) = \sigma^2$. Every time we query mother nature for Y at a given a x, she gives us $Y = h(x) + \epsilon$. A new ϵ is generated each time, independent of the last. We randomly sample some data $(x_i, y_i)_{i=1}^n$ and use it to fit a model $f_{\hat{\beta}}(x)$ according to some procedure (e.g. OLS, Ridge, LASSO). In class, we showed that

$$\underbrace{\mathbb{E}\left[(Y - f_{\hat{\beta}}(x))^2\right]}_{} = \underbrace{\sigma^2}_{} + \underbrace{(h(x) - \mathbb{E}\left[f_{\hat{\beta}}(x)\right])^2}_{} + \underbrace{\mathbb{E}\left[(\mathbb{E}\left[f_{\hat{\beta}}(x)\right] - f_{\hat{\beta}}(x))^2\right]}_{}.$$

(a) Label each of the terms above. Word bank: observation variance, model variance, observation bias^2, model bias^2, model risk, empirical mean square error.

Solution:

$$\underbrace{\mathbb{E}\left[(Y - f_{\hat{\beta}}(x))^2\right]}_{\text{model risk}} = \underbrace{\sigma^2}_{\text{observation variance}} + \underbrace{(h(x) - \mathbb{E}\left[f_{\hat{\beta}}(x)\right])^2}_{\text{model bias}^2} + \underbrace{\mathbb{E}\left[(\mathbb{E}\left[f_{\hat{\beta}}(x)\right] - f_{\hat{\beta}}(x))^2\right]}_{\text{model variance}}$$

(b) What is random in the equation above? Where does the randomness come from?

Solution: Y - this is the new observation at x. Its randomness comes from the noise generation process. $f_{\hat{\beta}}$ - this is the model fitted from the data. Its randomness comes from sampling and the noise generation process.

(c) True or false and explain. $\mathbb{E}\left[\epsilon f_{\hat{\beta}}(x)\right] = 0$

Solution: True. Since ϵ and $\hat{\beta}$ are independent,

$$\mathbb{E}\left[\epsilon f_{\hat{\beta}}(x)\right] = \mathbb{E}[\epsilon]\,\mathbb{E}\left[f_{\hat{\beta}}(x)\right] = 0$$

(d) Suppose you lived in a world where you could collect as many data sets you would like. Given a fixed algorithm to fit a model f_β to your data e.g. linear regression, describe a procedure to get good estimates of $\mathbb{E}\left[f_{\hat{\beta}}(x)\right]$ (technical point: you may assume this expectation exists).

Solution:

- Pick an x
- Gather a data set $\mathcal{D}_i$
- Fit a model $f_{\hat{\beta}_i}$ to that data set
- Calculate $f_{\hat{\beta}_i(x)}$
- Repeat many times
- Average over all the $f_{\hat{\beta}_i}(x)$

(e) If you could collect as many data sets as you would like, how does that affect the quality of your model $f_\beta(x)$?

Solution: By collecting many data sets, we have an unbiased estimate of the "average" model, but this does not mean our model will have unbiased prediction.

Ridge and LASSO Regression

2. Earlier, we posed the linear regression problem as follows: Find the $\vec{\beta}$ value that minimizes the average squared loss. In other words, our goal is to find $\hat{\vec{\beta}}$ that satisfies the equation below:

$$\hat{\vec{\beta}} = \underset{\vec{\beta}}{\text{argmin}}\, L(\vec{\beta}) = \underset{\vec{\beta}}{\text{argmin}}\, \frac{1}{n}||\vec{y} - \mathbb{X}\vec{\beta}||_2^2$$

Here, $\mathbb{X}$ is a $n \times d$ matrix, $\vec{\beta}$ is a $d \times 1$ vector and $\vec{y}$ is a $n \times 1$ vector. As we saw in lecture and in last week's discussion, the optimal $\hat{\vec{\beta}}$ is given by the closed form expression $\hat{\vec{\beta}} = (\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\vec{y}$.

To prevent overfitting, we saw that we can instead minimize the sum of the average squared loss plus a regularization function $\lambda\mathcal{S}(\vec{\beta})$. If use the function $\mathcal{S}(\vec{\beta}) = ||\vec{\beta}||_2^2$, we have "ridge regression". If we use the function $\mathcal{S}(\vec{\beta}) = ||\vec{\beta}||_1$, we have "LASSO regression". For example, if we choose S($\boldsymbol{\vec{\beta}}$) = $||\vec{\beta}||_2^2$, our goal is to find $\hat{\vec{\beta}}$ that satisfies the equation below:

$$\hat{\beta} = \underset{\vec{\beta}}{\text{argmin}}\, L(\vec{\beta}) = \underset{\vec{\beta}}{\text{argmin}}\, \frac{1}{n}||\vec{y} - \mathbb{X}\vec{\beta}||_2^2 + \lambda||\vec{\beta}||_2^2 = \underset{\vec{\beta}}{\text{argmin}}\, \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{X}_{i,\cdot}^T\vec{\beta})^2 + \lambda\sum_{j=1}^{d}\beta_j^2$$

Recall that λ is a hyperparameter that determines the impact of the regularization term. Though we did not discuss this in lecture, we can also find a closed form solution to ridge regression: $\hat{\vec{\beta}} = (\mathbb{X}^T\mathbb{X} + \lambda\mathbf{I})^{-1}\mathbb{X}^T\vec{y}$. It turns out that $\mathbb{X}^T\mathbb{X} + \lambda\mathbf{I}$ is guaranteed to be invertible (unlike $\mathbb{X}^T\mathbb{X}$ which might not be invertible).

(a) As model complexity increases, what happens to the bias and variance of the model?

Solution: Model complexity is inversely related to the regularization parameter λ. As λ increases, Bias tends to increase and variance tends to decrease.

(b) In terms of bias and variance, how does a regularized model compare to ordinary least squares regression?

Solution: Regularized regression has higher bias and lower variance relative to ordinary least squares regression.

(c) In ridge regression, what happens if we set $\lambda = 0$? What happens as λ approaches ∞?

Solution: If we set $\lambda = 0$ we end up with OLS. As λ approaches ∞ then $\boldsymbol{\beta}$ goes to $\mathbf{0}$.

(d) How does model complexity compare between ridge regression and ordinary least squares regression? How does this change for large and small values of λ?

Solution: Ridge regression in general will result in simpler models, as we penalize for large components in of $\boldsymbol{\beta}$. λ is inversely related to model complexity, e.g. larger values of λ represent larger penalties, meaning even lower model complexity.

(e) If we have a large number of features (10,000+) and we suspect that only a handful of features are useful, which type of regression (Lasso vs Ridge) would be more helpful in interpreting useful features?

Solution: LASSO would be better as it sets many values to 0, so it would be effectively selecting useful features and "ignoring" bad ones.

(f) What are the benefits of using ridge regression?

Solution: If $\mathbf{X}^T\mathbf{X}$ is not full rank (not invertible), then we end up with infinitely many solutions for least squares. But if we use ridge regression, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$. This guarantees invertbility and a unique solution, for $\lambda > 0$.

Cross Validation

3. After running 5-fold cross validation, we get the following mean squared errors for each fold and value of λ:

Fold Num	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	Row Avg
1	80.2	70.2	91.2	91.8	83.4
2	76.8	66.8	88.8	98.8	82.8
3	81.5	71.5	86.5	88.5	82.0
4	79.4	68.4	92.3	92.4	83.1
5	77.3	67.3	93.4	94.3	83.0
Col Avg	79.0	68.8	90.4	93.2	

How do we use the information above to choose our model? Do we pick a specific fold? a specific lambda? or a specific fold-lambda pair? Explain.

Solution: We should use $\lambda = 0.2$ because this value has the least average MSE across all folds.

4. You build a model with two regularization hyperparameters λ and γ. You have 4 good candidate values for λ and 3 possible values for γ, and you are wondering which λ, γ pair will be the best choice. If you were to perform five-fold cross-validation, how many validation errors would you need to calculate?

Solution: There are $4 \times 3 = 12$ pairs of λ, γ and each pair will have 5 validation errors, one for each fold.

5. In the typical setup of k-fold cross validation, we use a different parameter value on each fold, compute the mean squared error of each fold and choose the parameter whose fold has the lowest loss.

   ○ A. True

   ○ B. False