Business Analytics

Session 3b. Generalization and Bias-Variance Tradeoff

Renyu (Philip) Zhang

New York University Shanghai

Spring 2019

Generalization

Population v.s. Sample

- We have a *sample* data $\mathcal{D} = \{X_{ij}, Y_i : 1 \le i \le n, 1 \le j \le p\}$.
- The *sample* data comes from a *population*.
  - Think about surveys: Understand the broader population through a (much) smaller sample.
- Both the sample and the population comes from some probablistic data generating process.
- Generalization: Use the sample to reason about the data generate process and the population.

Regression and Classification in a Unified Framework

- **Goal**: Given data $\mathcal{D} = \{X_{ij}, Y_i : 1 \le i \le n, 1 \le j \le p\}$, fit a model $\hat{f}(\cdot)$, such that

  $$\text{the generalization error } \mathbb{E}[\mathcal{L}(Y - \hat{f}(X))] \text{ is minimized,}$$

  where $\mathcal{L}(\cdot)$ is a loss/error function, and the expectation is taken with respect to the distribution that generates the data.

- Loss function $\mathcal{L}(\cdot)$: $(Y - \hat{f}(X))^2$, $|Y - \hat{f}(X)|$, $\mathbf{1}\{Y \neq \hat{f}(X)\}$, etc.

Train-Validate-Test

1. Separate data into three groups: *training, validation and testing.*

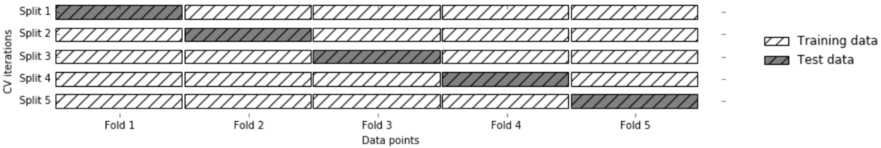
2. *Training*: Use training data to build different candidates of models $\hat{f}_1(\cdot), \hat{f}_2(\cdot), \ldots, \hat{f}_L(\cdot)$.

3. *Validation*: Use validation data to estimate the generalization error $\mathbb{E}[\mathcal{L}(Y - \hat{f}_i(X))]$ of each model i, and pick up the best one with the smallest generalization error, $\hat{f}_*(\cdot)$.

4. *Testing*: Use the tesing data to assess the performance of the chosen model $\hat{f}_*(\cdot)$.

The estimated error on the testing data is an unbiased estimation of the generalization error.

$k-$fold Cross Validation



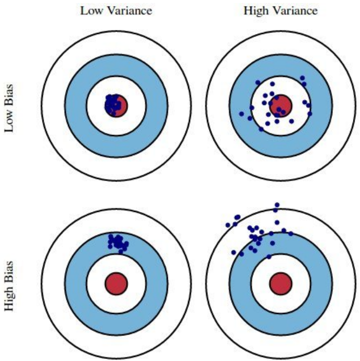
- Alternative approach: $k-$fold cross validation (CV).
- CV is more stable (with smaller variance), requires fewer data, and slower than the train-validation-test approach.
- In practice, we often use k=5 to 10.
- Stratification: Relative class frequencies in each fold reflect relative class frequencies on the whole dataset.

Bias-Variance Decomposition

- Assume the data (X_i, Y_i) are generated according to $Y_i = f(X_i) + \epsilon_i$ where $\mathbb{E}(\epsilon_i) = 0$. We want to find $\hat{f}(\cdot)$ (from data) so that $Y_i \approx \hat{f}(X_i)$ and the error is as small as possible.
- Bias-Variance decomposition with squared error:

$$\underbrace{\mathbb{E}(Y_i - \hat{f}(X_i))^2}_{\text{Squared Error}} = \underbrace{Var(\hat{f}(X_i))}_{\text{Variance}} + \underbrace{\mathbb{E}(\hat{f}(X_i) - f(X_i))^2}_{\text{Bias}} + \underbrace{Var(Y_i)}_{\text{Noise}}$$

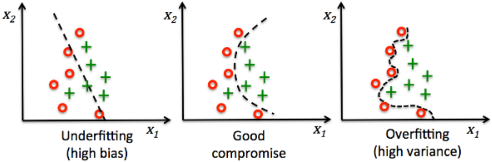
- Variance: How much your model will change if you train on a different data set? Measures overfitting.
- Bias: What is the inherent error with your *model* even if you have infinitely many training data? Measures underfitting.
- Noise: How big is the intrinsic noise?

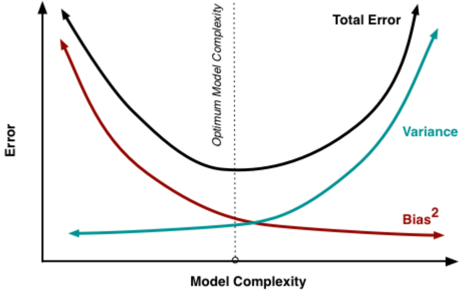


Bias-Variance Tradeoff

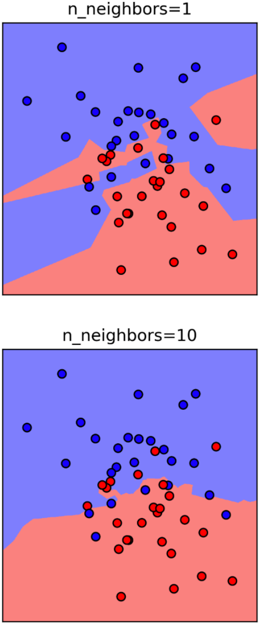
- More "complex" model overfits the training data, variance $\uparrow$ and bias $\downarrow$.

- Less "complex" model underfits the training data, variance $\downarrow$ and bias $\uparrow$.





k–NN for Different Number of Neighbors



Addressing Over-fitting: Regularization

- Regularization means to penalize overly complex models.
  - The fitted model will not over fit the training data.
- Examples:

$$\text{Lasso linear regression: } \min_{\hat{\beta}} SSE + \alpha \sum_{j=1}^{p} |\hat{\beta}_j|$$

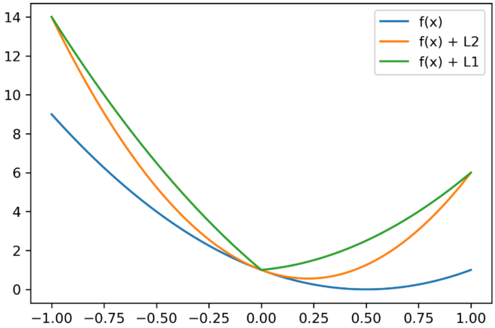
$$\text{Ridge linear regression: } \min_{\hat{\beta}} SSE + \alpha \sum_{j=1}^{p} |\hat{\beta}_j|^2$$

$$\text{Lasso logistic regression: } \max_{\hat{\beta}} \text{Log-Likelihood} + \alpha \sum_{j=1}^{p} |\hat{\beta}_j|$$

$$\text{Ridge logistic regression: } \max_{\hat{\beta}} \text{Log-Likelihood} + \alpha \sum_{j=1}^{p} |\hat{\beta}_j|^2$$

- The parameter $\alpha \geq 0$ trades off bias (under-fitting) and variance (over-fitting).
  - The larger the α, the more we penalize over-fitting.

Lasso (L_1) and Ridge (L_2) Regularizations



$$f(x) = (2x-1)^2$$
$$f(x) + L2 = (2x-1)^2 + \alpha x^2$$
$$f(x) + L1 = (2x-1)^2 + \alpha|x|$$

- Lasso and Ridge regressions will yield coefficients $\hat{\beta}$ that "shrink" to zero.
- The most explanatory covariates will be retained.
- Lasso typically yields a much smaller subset of nonzero coefficients than ridge or OLS (i.e., fewer nonzero entries in $\hat{\beta}$).
- Use train-valid-test or cross validation to tune the parameter α.

Homework

- Finish Homework 3 (NO need to submit it).

- Read "*The Analytics Edge*", Chapters 8.