Statistical Learning - Ch 1,2 (Oct 13, 2018 10 am)
Introduction and Overview to Statistical Learning

Discussions

- **What is Statistical Learning? Or what is Statistics?**
  From the course, statistical learning is a subset of statistics. The related field of machine learning is a subset of AI. There is a big intersection between statistical learning and machine learning. The main difference between statistical learning and machine learning is that
- **What is the Bayes classifier and its relationship to K Nearest Neighbors?**

  $$C^{\text{Bayes}}(x) = \underset{r \in \{1,2,\ldots,K\}}{\text{argmax}} \; \text{P}(Y = r \mid X = x).$$

  The Bayes classifier is the optimal classification provided some input data: choose the class with maximal probability given the input. K nearest neighbors is a practical way of approximating this optimal classifier. Wikipedia puts it nicely: "An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors." K-NN effectively uses a sample of similar objects to approximate the true (population) conditional probability distribution that appears in the Bayes classifer.
  https://en.wikipedia.org/wiki/Bayes_classifier
  https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
- **Intuitively, why is nearest neighbor lousy when the number of parameters p (the dimensionality) is large?**
  It is hard to visualize high dimensional space and so high dimensional space is not intuitive. The informal reason why the nearest neighbor algorithm is lousy in high dimensional space is that points in high dimensional space are more spread out.
- **What does it mean to stay local?**
  It means to prefer products made in Hawaii rather than products shipped in to Hawaii.
- **What is overfitting vs. underfitting vs. a good fit in relation to the bias-variance tradeoff?**

  $$E\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon). \qquad (2.7)$$

  - Overfitting = high variance
  - Underfitting = high bias
  - Good fit = low variance, low bias
  - Best fit = minimum variance + bias squared

  Note: Only the MSE can be decomposed into variance + bias^2. Other loss functions such as log loss or mean absolute error don't have this nice decomposition. Nevertheless, the bias-variance tradeoff is a useful tool in understanding how complexity of a model affects the predictions (high complexity leads to high variance and low bias, low complexity leads to low variance and high bias).
- **What is a statistical model?**
  y_hat = f(X; theta) f is our statistical model. More concretely, we have a statistical model if y = f(X; theta) + epsilon, where epsilon is some random error term; that is, we don't ever have the true mapping between X and y, instead we strive for a mapping y_hat = f(X; theta) such that y_hat is close to y. We can contrast statistical model with scientific model which is derived from first principles.