Machine Learning – Brett Bernstein

Week 2 Lecture: Concept Check Exercises

Starred problems are optional.

Excess Risk Decomposition

1. Let $\mathcal{X} = \mathcal{Y} = \{1, 2, \ldots, 10\}$, $\mathcal{A} = \{1, \ldots, 10, 11\}$ and suppose the data distribution has marginal distribution $X \sim \text{Unif}\{1, \ldots, 10\}$. Furthermore, assume $Y = X$ (i.e., Y always has the exact same value as X). In the questions below we use square loss function $\ell(a, x) = (a - x)^2$.

   (a) What is the Bayes risk?

   (b) What is the approximation error when using the hypothesis space of constant functions?

   (c) Suppose we use the hypothesis space $\mathcal{F}$ of affine functions.

      i. What is the approximation error?

      ii. Consider the function $\hat{f}(x) = x + 1$. Compute $R(\hat{f}) - R(f_{\mathcal{F}})$.

2. ($\star$) Let $\mathcal{X} = [-10, 10]$, $\mathcal{Y} = \mathcal{A} = \mathbb{R}$ and suppose the data distribution has marginal distribution $X \sim \text{Unif}(-10, 10)$ and $Y|X = x \sim \mathcal{N}(a + bx, 1)$. Throughout we assume the square loss function $\ell(a, x) = (a - x)^2$.

   (a) What is the Bayes risk?

   (b) What is the approximation error when using the hypothesis space of constant functions (in terms of a and b)?

   (c) Suppose we use the hypothesis space of affine functions.

      i. What is the approximation error?

      ii. Suppose you have a fixed data set and compute the empirical risk minimizer $\hat{f}_n(x) = c + dx$. What is the estimation error (int terms of a, b, c, d) ?

3. Try to best characterize each of the following in terms of one or more of optimization error, approximation error, and estimation error.

   (a) Overfitting.

   (b) Underfitting.

   (c) Precise empirical risk minimization for your hypothesis space is computationally intractable.

   (d) Not enough data.

4. (a) We sometimes look at $R(\hat{f}_n)$ as random, and other times as deterministic. What causes this difference?

   (b) True or False: Increasing the size of our hypothesis space can shift risk from approximation error to estimation error but always leaves the quantity $R(\hat{f}_n) - R(f^*)$ constant.

   (c) True or False: Assume we treat our data set as a random sample and not a fixed quantity. Then the estimation error and the approximation error are random and not deterministic.

   (d) True or False: The empirical risk of the ERM, $\hat{R}(\hat{f}_n)$, is an unbiased estimator of the risk of the ERM $R(\hat{f}_n)$.

   (e) In each of the following situations, there is an implicit sample space in which the given expectation is computed. Give that space.

      i. When we say the empirical risk $\hat{R}(f)$ is an unbiased estimator of the risk $R(f)$ (where f is independent of the training data used to compute the empirical risk).

      ii. When we compute the expected empirical risk $\mathbb{E}[R(\hat{f}_n)]$ (i.e., the outer expectation).

      iii. When we say the minibatch gradient is an unbiased estimator of the full training set gradient.

5. For each, use $\leq$, $\geq$, or $=$ to determine the relationship between the two quantities, or if the relationship cannot be determined. Throughout assume $\mathcal{F}_1, \mathcal{F}_2$ are hypothesis spaces with $\mathcal{F}_1 \subseteq \mathcal{F}_2$, and assume we are working with a fixed loss function ℓ.

   (a) The estimation errors of two decision functions f_1, f_2 that minimize the empirical risk over the same hypothesis space, where f_2 uses 5 extra data points.

   (b) The approximation errors of the two decision functions f_1, f_2 that minimize risk with respect to $\mathcal{F}_1, \mathcal{F}_2$, respectively (i.e., $f_1 = f_{\mathcal{F}_1}$ and $f_2 = f_{\mathcal{F}_2}$).

   (c) The empirical risks of two decision functions f_1, f_2 that minimize the empirical risk over $\mathcal{F}_1, \mathcal{F}_2$, respectively. Both use the same fixed training data.

   (d) The estimation errors (for $\mathcal{F}_1, \mathcal{F}_2$, respectively) of two decision functions f_1, f_2 that minimize the empirical risk over $\mathcal{F}_1, \mathcal{F}_2$, respectively.

   (e) The risk of two decision functions f_1, f_2 that minimize the empirical risk over $\mathcal{F}_1, \mathcal{F}_2$, respectively.

6. In the excess risk decomposition lecture, we introduced the decision tree classifier spaces $\mathcal{F}$ (space of all decision trees) and $\mathcal{F}_d$ (the space of decision trees of depth d) and went through some examples. The following questions are based on those slides. Recall that $P_{\mathcal{X}} = \text{Unif}([0,1]^2)$, $\mathcal{Y} = \{\text{blue}, \text{orange}\}$, orange occurs with .9 probability below the line $y = x$ and blue occurs with .9 probability above the line $y = x$.

(a) Prove that the Bayes error rate is 0.1.

(b) Is the Bayes decision function in $\mathcal{F}$?

(c) For the hypothesis space $\mathcal{F}_3$ the slide states that $R(\tilde{f}) = 0.176 \pm .004$ for $n = 1024$. Assuming you had access to the training code that produces $\tilde{f}$ from a set of data points, and random draws from the data generating distribution, give an algorithm (pseudocode) to compute (or estimate) the values 0.176 and .004.

L_1 and L_2 Regularization

1. Consider the following two minimization problems:

$$\arg\min_w \Omega(w) + \frac{\lambda}{n}\sum_{i=1}^{n} L(f_w(x_i), y_i)$$

and

$$\arg\min_w C\Omega(w) + \frac{1}{n}\sum_{i=1}^{n} L(f_w(x_i), y_i),$$

where $\Omega(w)$ is the penalty function (for regularization) and L is the loss function. Give sufficient conditions under which these two give the same minimizer.

2. ($\star$) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Prove that $\|\nabla f(x)\|_2 \leq L$ if and only if f is Lipschitz with constant L.

3. ($\star$) Let $\hat{w}$ denote the minimizer for

$$\begin{array}{ll} \text{minimize}_w & \|Xw - y\|_2^2 \\ \text{subject to} & \|w\|_1 \leq r. \end{array}$$

Prove that $f(x) = \hat{w}^T x$ is Lipschitz with constant r.

4. Two of the plots in the lecture slides use the fact that $\|\hat{\beta}\|/\|\tilde{\beta}\|$ is always between 0 and 1. Here $\hat{\beta}$ is the parameter vector of the linear model resulting from the regularized least squares problem. Analgously, $\tilde{\beta}$ is the parameter vector from the unregularized problem. Why is this true that the quotient lies in $[0, 1]$?

5. Explain why feature normalization is important if you are using L_1 or L_2 regularization.