NYU Center for Data Science: DS-GA 1003

Machine Learning and Computational Statistics (Spring 2019)

Brett Bernstein

February 4, 2019

Instructions: Following most lab and lecture sections, we will be providing concept checks for review. Each concept check will:

- List the lab/lecture learning objectives. You will be responsible for mastering these objectives, and demonstrating mastery through homework assignments, exams (midterm and final), and on the final course project.
- Include concept check questions. These questions are intended to reinforce the lab/lectures, and help you master the learning objectives.

You are strongly encourage to complete all concept check questions, and to discuss these (and related) problems on Piazza and at office hours. However, problems marked with a ($\star$) are considered optional.

Lecture 2: Excess Risk Decomposition, Stochastic Gradient Descent and Regularization

Topic 1: Excess Risk Decomposition

Learning Objectives

1. Give precise definitions for excess risk, approximation error, estimation error, and optimization error.
2. Suppose we have nested hypothesis spaces, say $\mathcal{H}_1 \subset \mathcal{H}_2$. Explain how we would expect the approximation error and estimation error to change when we change from $\mathcal{H}_1$ to $\mathcal{H}_2$, all else fixed.
3. Explain how we would expect the approximation error and estimation error to change when we increase the sample size, all else fixed.
4. Explain optimization error, and write down an excess risk decomposition that incorporates approximation error, estimation error, and optimization error. Why might we have negative optimization error but never negative estimation error?

Concept Check Questions

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.

Topic 2: Stochastic Gradient Descent

Learning Objectives

1. Be able to write the empirical risk for a particular loss function over a particular parameterized hypothesis space, such as for square loss over a hypothesis space of linear functions.

2. Compare and constrast gradient descent, minibatch gradient descent, and stochastic gradient descent.

Concept Check Questions

1. When performing mini-batch gradient descent, we often randomly choose the mini-batch from the full training set without replacement. Show that the resulting mini-batch gradient is an unbiased estimate of the gradient of the full training set. Here we assume each decision function f_w in our hypothesis space is determined by a parameter vector $w \in \mathbb{R}^d$.

2. You want to estimate the average age of the people visiting your website. Over a fixed week we will receive a total of N visitors (which we will call our full population). Suppose the population mean μ is unknown but the variance σ^2 is known. Since we don't want to bother every visitor, we will ask a small sample what their ages are. How many visitors must we randomly sample so that our estimator $\hat{\mu}$ has variance at most $\epsilon > 0$?

3. ($\star$) Suppose you have been successfully running mini-batch gradient descent with a full training set size of 10^5 and a mini-batch size of 100. After receiving more data your full training set size increases to 10^9. Give a heuristic argument as to why the mini-batch size need not increase even though we have 10000 times more data.

Topic 3: L_1 and L_2 Regularization

Learning Objectives

1. Explain the concept of a sequence of nested hypothesis spaces, and explain how a complexity measure (of a function) can be used to create such a sequence.

2. Given a base hypothesis space of decision functions (e.g. affine functions), a performance measure for a decision function (e.g. empirical risk on a training set), and a function complexity measure (e.g. Lipschitz continuity constant of decision function), give the corresponding optimization problem in Tikhonov and Ivanov forms.

3. For some situations (i.e. combinations of base hypothesis space, performance measure, and complexity measure), we claimed that Tikhonov and Ivanov forms are equivalent. Be able to explain what this means and write it down mathematically.

4. In particular, the Tikhonov and Ivanov formulations are equivalent for lasso and ridge regression. Be comfortable switching between the formulations to assist with interpretations (e.g. the classic L1 regularization picture with the norm ball is based on the Ivanov formulation).

Concept Check Questions

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 $\|\hat{w}\|/\|\tilde{w}\|$ is always between 0 and 1. Here $\hat{w}$ is the parameter vector of the linear model resulting from the regularized least squares problem. Analgously, $\tilde{w}$ 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.