Foundations of Machine Learning

Brett Bernstein

August 22, 2018

Lecture 2: Excess Risk Decomposition 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 $\mathcal{X} = \mathcal{Y} = \{1, 2, \dots, 10\}$, $\mathcal{A} = \{1, \dots, 10, 11\}$ and suppose the data distribution has marginal distribution $X \sim \text{Unif}\{1, \dots, 10\}$. 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 $R(\hat{f}) - R(f_{\mathcal{F}})$.

Solution.

(a) The best decision function is $f^*(x) = x$. The associated risk is 0.

(b) The best constant function is $f(x) = \mathbb{E}[Y] = \mathbb{E}[X] = 5.5$. This has risk

$$\mathbb{E}[(Y - 5.5)^2] = \mathrm{Var}(Y) = \frac{33}{4},$$

by using (or deriving) the formula for the variance of a discrete uniform distribution. Thus the approximation error is $33/4$.

(c) i. The Bayes decision function is affine, so the approximation error is 0.

ii. The risk is

$$R(\hat{f}) = \mathbb{E}[(Y - \hat{f}(X))^2] = \mathbb{E}[(X - (X + 1))^2] = 1.$$

Thus the answer is 1.

2. ($\star$) Let $\mathcal{X} = [-10, 10]$, $\mathcal{Y} = \mathcal{A} = \mathbb{R}$ and suppose the data distribution has marginal distribution $X \sim \mathrm{Unif}(-10, 10)$ 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) ?

Solution. Throughout we use the fact that $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

(a) The best decision function is $f(x) = \mathbb{E}[Y|X = x] = a + bx$. This has risk

$$\mathbb{E}[(Y - a - bX)^2] = \mathbb{E}[\mathbb{E}[(Y - a - bX)^2|X]] = \mathbb{E}[1] = 1.$$

(b) The best constant function is given by $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = a + b\mathbb{E}[X] = a$. This has risk

$$\mathbb{E}[(Y - a)^2] = \mathbb{E}[\mathbb{E}[(Y - a)^2|X]] = \mathbb{E}[1 + b^2X^2] = 1 + b^2\mathbb{E}[X^2],$$

where

$$\mathbb{E}[X^2] = \int_{-10}^{10} \frac{x^2}{20}\, dx = \frac{2000}{3 \cdot 20} = \frac{100}{3}.$$

Thus the approximation error is $100b^2/3$.

(c) i. There is an affine Bayes decision function, so the approximation error is 0.

ii. Note that

$$\begin{aligned} R(\hat{f}_n) &= \mathbb{E}[(Y - c - dX)^2] = \mathbb{E}[\mathbb{E}[(Y - c - dX)^2 | X]] \\ &= \mathbb{E}[1 + ((a - c) + (b - d)X)^2] = 1 + (a - c)^2 + 100(b - d)^2/3. \end{aligned}$$

Thus the estimation error is $(a - c)^2 + 100(b - d)^2/3$.

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.

Solution.

(a) High estimation error due to insufficient data relative to the complexity of your hypothesis space. Can be accompanied by low approximation error indicating a complex hypothesis space.

(b) High approximation error due to an overly simplistic hypothesis space. Can be accompanied by low estimation estimation error due to the large amount of data relative to the (low) complexity of the hypothesis space.

(c) Increased optimization error.

(d) High estimation error.

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.

Solution.

(a) The quantity is random when we consider the training data as a random sample of size n. If we focus on a fixed set of training data then the quantity is deterministic.

(b) False. Note that $\hat{f}_n$ depends on which hypothesis space you have chosen. As an example, imagine having an affine Bayes decision function, and changing the hypothesis space from the set of affine functions to the set of all decision functions. This can cause empirical risk minimization to overfit the training data thus creating a sharp rise in $R(\hat{f}_n) - R(f^*)$.

(c) False, approximation error is a deterministic quantity.

(d) False. The empirical risk of the ERM will often be biased low. This is why we use a test set to approximate its true risk. The issue is that $\hat{f}_n$ depends on the training data so

$$\mathbb{E}\ell(\hat{f}_n(x_i), y_i) \neq \mathbb{E}\ell(\hat{f}_n(x), y)$$

where x, y is a new random draw from the data distribution that isn't in the training data.

(e) i. The space of training sets (i.e., samples of size n from the data generating distribution).

ii. The space of training sets (i.e., samples of size n from the data generating distribution).

iii. The space of all minibatches chosen from the full training set (i.e., samples of of the batch size from the empirical distribution on the full training set).

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.

Solution.

(a) Roughly speaking, more data is better, so we would tend to expect that f_2 will have lower estimation error. That said, this is not always the case, so the relationship cannot be determined.

(b) The approximation error of f_1 will be larger.

(c) The empirical risk of f_1 will be larger.

(d) Roughly speaking, increasing the hypothesis space should increase the estimation error since the approximation error will decrease, and we expect to need more data. That said, this is not always the case, so the answer is the relationship cannot be determined.

(e) Cannot be determined.

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.

Solution.

(a) Since the output space is discrete and we are using the $0-1$ loss, our best prediction is the highest probability output conditional on the input. By choosing orange below the line $y = x$ and blue above, we obtain a .1 probability of error. For the $0-1$ loss, probability of error gives the risk.

(b) No. Any decision tree in $\mathcal{F}$ has finite depth, and thus will divide $[0,1]^2$ into a finite number of rectangles. Thus we cannot produce the decision boundary $y = x$ used by the Bayes decision function.

(c) Pseudocode follows:

i. Initialize L to be an empty list of risks.

ii. Repeat the following M times for some sufficiently large M:

   A. Draw a random sample $(x_1, y_1), \ldots, (x_n, y_n)$ from the data generating distribution.

   B. Obtain a decision function $\tilde{f}$ by running our training algorithm on the generated sample.

   C. Draw a new random sample $(x'_1, y'_1), \ldots, (x'_S, y'_S)$ of size S where S is sufficiently large.

   D. Compute $e = |\{i \mid \tilde{f}(x'_i) \neq y'_i\}|$. That is, the number of times $\tilde{f}$ is incorrect on our new sample.

   E. Add e/S to the list L.

iii. Compute the sample average and standard deviation of the values in L. Above .176 would be the average and .004 would be the standard deviation.

Instead of drawing the sample of size S we could have computed the risk analytically.

Topic 2: 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.

Solution. Let $C = 1/\lambda$. Then the two objectives differ by a constant factor.

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.

   Solution. First suppose $\|\nabla f(x)\|_2 \leq L$ for some $L \geq 0$ and all $x \in \mathbb{R}^n$. By the mean value theorem we have, for any $x, y \in \mathbb{R}^n$,

   $$f(y) - f(x) = \nabla f(x + \xi(y - x))^T (y - x),$$

   where ξ is some value between 0 and 1. Taking absolute values on each side we have

   $$|f(y) - f(x)| = |\nabla f(x + \xi(y - x))^T (y - x)| \leq \|\nabla f(x + \xi(y - x))\|_2 \|y - x\|_2$$

   by Cauchy-Schwarz. Applying our bound on the gradient norm proves f is Lipschitz with constant L.

   Conversely, suppose f is Lipschitz with constant L. Note that

   $$|\nabla f(x)^T v| = |f'(x; v)| = \left|\lim_{t \to 0} \frac{f(x + tv) - f(x)}{t}\right| \leq \lim_{t \to 0} \frac{|t| L \|v\|}{|t|} = L\|v\|.$$

   Letting $v = \nabla f(x)$ we obtain $\|\nabla f(x)\|_2^2 \leq L\|\nabla f(x)\|_2$ giving the result.

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.

   Solution. Note that $\|w\|_2 \leq \|w\|_1 \leq r$, so the argument from class gives the result. To see the inequality, note that

   $$\|w\|_1^2 = (|w_1| + \cdots + |w_n|)^2 \geq |w_1|^2 + \cdots + |w_n|^2 = \|w\|_2^2.$$

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 $[0, 1]$?

Solution. We assume Ivanov regularization (since Tikhonov is equivalent). We know that

$$\frac{1}{n}\sum_{i=1}^{n}(\tilde{w}^T x_i - y_i)^2 \leq \frac{1}{n}\sum_{i=1}^{n}(\hat{w}^T x_i - y_i)^2$$

since $\tilde{w}$ is the solution to the unconstrained minimization. But if $\|\tilde{w}\| \leq \|\hat{w}\|$ then $\|\tilde{w}\|$ is feasible for the regularized problem, so $\|\hat{w}\| = \|\tilde{w}\|$. Thus $\|\tilde{w}\| \geq \|\hat{w}\|$.

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

   Solution. Suppose you have a model $y = w^T x$ where x_1 is a very correlated with y, but the feature is measured in meters. Thus $w_1 = 4$ would mean each increase in x_1 by 1 meter yields an increase in y by 4. Now suppose we change the units of w_1 to kilometers by scaling it. This would require us to change w_1 to 4000 to achieve the same decision function. While this has no effect on the loss $(y - w^T x)^2$ it has a significant effect on $\lambda\|w\|_2^2$ or $\lambda\|w\|_1$. For example, even if $x_2, \ldots, x_n$ had very little relationship with y, we would still undervalue w_1 due to the regularization.