Section 03: Bias-Variance Tradeoff and Ridge Regression

1. Bias-Variance Trade-off

Consider a simple statistical learning setting, in which we assume that there is some unknown function relating two random variables X and Y (e.g. $Y = 2X$). Let us denote this function by $Y = \eta(X)$; however, we don't know specifically what this function $\eta(\cdot)$ is. Our goal is as follows. Given X, we want to predict Y with the smallest possible error, in expectation. We formalize this notion below.

(a) Find the function η that minimizes the expected squared error $\mathbb{E}[(Y - \eta(X))^2]$. **Hint:** Observe from problem 2a of HW 0 that $\mathbb{E}[(Y - \eta(X))^2] = \mathbb{E}[\mathbb{E}[(Y - \eta(X))^2|X = x]]$ (The "Tower Rule").

(b) While ideally we want η to be what we computed above, in reality, however, we are restricted to our training data and a function class, the best we can do is
$\hat{f}_D = \arg\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$, where $D = \{(x_i, y_i)\}$. Here, (x_i, y_i) is a sample from distribution P_{XY}. To account for the prediction error (i.e. quality of our estimator $\hat{f}_D$), we need to calculate

$$E[E_D[(Y - \hat{f}_D(x))^2]|X = x]]$$

We can break the expectation into

$$\mathbb{E}[\mathbb{E}[(Y - \eta(x))^2|X = x]] + \mathbb{E}_D[(\eta(x) - \hat{f}_D(x))^2]$$

$\mathbb{E}[\mathbb{E}[(Y - \eta(x))^2|X = x]]$ is called **irreducible error** — the error incurred even in ideal situation.

$\mathbb{E}_D[(\eta(x) - \hat{f}_D(x))^2]$ is called **learning error** — the error incurred by the learning setting (e.g. insufficient data, the chosen model class F is not expressive enough etc.)

Express the **learning error** in terms of

- bias — $(\eta(x) - \mathbb{E}_D[\hat{f}_D(x)])$
- and variance — $\mathbb{E}_D[(\mathbb{E}_D[\hat{f}_D(x)] - \hat{f}_D(x))^2]$

and explain why there is a trade-off.

2. A General View of Regularized Least Squares Regression

We saw linear regression as well as ridge regression in class. Here we consider a problem that generalizes both of these. As a reminder, in linear regression, we seek a model that captures a linear relationship between input data and output data. The general case we consider imposes additional structure on the model.

Consider an experiment in which you have n data points $x_i \in \mathbb{R}^d$ and corresponding n observations y_i. We wish to come up with a model $\omega \in \mathbb{R}^d$ that satisfies the following properties: first, the error $\sum_{i=1}^n (x_i^\top \omega - y_i)^2$ should be small; second, we don't want small changes in training data resulting in large changes in solution; third, we want to put different weights in controlling the magnitude of different coordinates of ω. We therefore define

$$\widehat{\omega}_{\text{general}} = \arg\min_{\omega} \sum_{i=1}^{n} (y_i - x_i^\top \omega)^2 + \lambda \sum_{i=1}^{d} D_{ii} \omega_i^2.$$

Here, D is a diagonal matrix, with positive entries on the diagonal. Observe that when D is the identity matrix, we recover ridge regression, and when $\lambda = 0$, we recover least squares regression. Different weights on D_{ii} cause the magnitudes of ω_i to be controlled differently.

2.1. Closed form in the general case

Deduce the closed form solution for $\widehat{\omega}_{\text{general}}$. You should be comfortable with proofs in the "coordinate" form as well as the "matrix" form.

2.2. Special cases: linear regression and ridge regression

(a) In the simple least squares case ($\lambda = 0$ above), what happens to the resulting $\widehat{\omega}$ if we double all the values of y_i?

(b) In the simple least squares case ($\lambda = 0$ above), what happens to the resulting $\widehat{\omega}$ if we double the data matrix $X \in \mathbb{R}^{n \times d}$?

(c) Suppose $D = I$ (that is, it is the identity matrix). That is, this is the *ridge* regression setting. Explain why $\lambda > 0$ ensures a "well-conditioned" setting (by "well-conditioned" we mean that matrix in the solution is positive definite and thus invertible).

3. Understanding Stein's Paradox through bias-variance trade-off

In this problem, we'll use bias-variance tradeoff to find a non-obvious way of estimating the mean of unrelated distributions.

So far in class, we've always been trying to learn a function – given a bunch of features, understand how they predict the single-number output. In this problem, we're trying to do something a little different. We have n completely unrelated probability distributions. We're going to get one sample from each of the distributions, and attempt to predict each of their means. For some examples, our distributions might be: high temperature in Chicago on January 1st, low temperature in Seattle on December 1st, and your friend's score on the midterm.

More formally, let $\theta \in \mathbb{R}^n$ be the (unknown) true means of our n distributions. We will get a vector X where each $X_i \sim \mathcal{N}(\theta_i, \sigma^2)$. We're assuming that every distribution has the same variance, but our means could be very different. Our job is to report $\hat{\theta}$ to minimize our expected error: $\mathbb{E}\left[||\hat{\theta} - \theta||_2^2\right]$.

3.1. The Maximum Likelihood Estimator

The most natural estimator is the maximum likelihood estimator $\hat{\theta} = X$. It's not obvious that any other viable strategy exists. We'll use bias-variance tradeoff to show that there's actually a better estimator.

(a) Split the error into bias^2 and variance. I.e. show

$$\mathbb{E}[||\hat{\theta} - \theta||_2^2] = ||\mathbb{E}[\hat{\theta}] - \theta||_2^2 + \mathbb{E}[||\hat{\theta} - \mathbb{E}[\hat{\theta}]||_2^2]$$

Hint: add and subtract $\mathbb{E}[\hat{\theta}]$.

(b) What is the variance of the estimator $\hat{\theta} = X$? Hint: Remember that for a random variable Z, $\text{Var}(Z) = \mathbb{E}[(Z - \mathbb{E}[Z])^2]$

(c) What is the bias^2 of the estimator $\hat{\theta} = X$?

3.2. A Biased Estimator

The maximum likelihood estimator above is an unbiased estimator. However, if our goal is to minimize the expected mean squared error, can we sacrifice some bias to lower variance dramatically. Here's an estimator to consider: we shrink/bias the MLE estimate towards the mean of all observations, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. To be exact, we would like to consider the estimator $\hat{\theta} = (1-\lambda)X + \lambda\overline{X}\mathbf{1}$ for some $0 \leq \lambda \leq 1$ where $\mathbf{1}$ is a vector will all 1s. We will show that for any θ, there will always be a such estimator (by choosing a proper λ) that better minimizes the mean squared error than the MLE. But this should sound crazy in two ways:

- We're intentionally guessing something we *know* is a biased estimator;
- **More importantly, now when we estimate the value $\hat{\theta}$, it becomes $(1-\lambda)X + \lambda\overline{X}\mathbf{1}$, where $\overline{X}$ depends on all observations $X_1, ..., X_n$.** Remember $X_1, ..., X_n$ are independent random variables.

(a) What is the variance of the estimator $\hat{\theta} = (1-\lambda)X + \lambda\overline{X}\mathbf{1}$?

(b) What is the bias^2 of the estimator?

(c) What value of λ will result in the best estimator?

(d) Compare the error you get from this biased estimator with unbiased estimator you have from section 2.1. Which one has smaller error?

This is what Stein's paradox is pointing at — **we might need to intentionally inject bias, to reduce the overall error.** Maybe the injected bias uses seemingly irrelevant information (e.g. other independent variables).

Further Reading: James-Stein Estimator

The Bias-Variance Tradeoff says that since our error is just the sum of the bias^2 and the variance, if we can find a way to "tradeoff" bias for variance, we can affect our error. With our previous estimator, the two sources of error are quite imbalanced. None of our error is from bias, it all comes from variance. Can we think of a way to reduce variance (even if it means increasing the bias)?

Normally, the way we would reduce variance would be to sample the random variables again and take the average of the samples. But we can't do that for this problem (it would take us a whole year to get another high temperature on January 1st). Another way to decrease the variance is to "scale down" the random variable.

For example, in this section we are going to use the "James-Stein Estimator", $\hat{\theta}^{JS} = \left(1 - \frac{(n-2)}{||X||_2^2}\right) X$, where we scale our estimate by a factor of $1 - \frac{(n-2)}{||X||_2^2}$ less than 1.

It has been shown in [1], for any θ, the expected mean squared error of $\hat{\theta}^{JS}$ is consistently better than $\hat{\theta}^{MLE}$'s, i.e.,

$$\mathbb{E}[||\hat{\theta}^{JS} - \theta||_2^2] \leq \mathbb{E}[||\hat{\theta}^{MLE} - \theta||_2^2], \qquad \forall \theta \in \mathbb{R}^n$$

where $\hat{\theta}^{MLE} = X$ is the maximum likelihood estimator from 2.1. **Our estimator doesn't even depend on obtaining the optimal hyperparamter λ (as in 2.2) anymore.**

To understand the paradox, here are some essential ingredients and outline:

- The squared error is the sum of n individual errors — $\mathbb{E}[(\theta_1 - \hat{\theta}_1)^2], \mathbb{E}[(\theta_2 - \hat{\theta}_2)^2], ...$ In this case, the errors are calculated by squaring deviation from the true mean.
- When we do the scaling proposed by James and Stein, some of these individual errors will decrease, while some will increase.
- However, notice the function $(\cdot)^2$ will penalize the higher deviation more harshly than the smaller deviation, even though the total deviation remains the same.
- For example, $\mathbb{E}[(\theta_1 - \hat{\theta}_1)^2]$ could change from 100^2 to 90^2 while $\mathbb{E}[(\theta_2 - \hat{\theta}_2)]$ changes from 0 to 10^2, the overall squared error will decrease by $(100^2 - 90^2) - (0^2 - 10^2) = 1906$!

Now that we know how $\hat{\theta}^{JS}$ is trying to improve the overall squared error, we see that it is the objective $\min_{\hat{\theta}} \mathbb{E}[||\hat{\theta} - \theta||_2^2] = \sum_{i=1}^{n} \mathbb{E}[(\hat{\theta}_i - \theta_i)^2]$ that is dependent on all of the random variables $X_1, ..., X_n$. And to minimize this overall/joint objective, a good estimator of θ_i necessarily becomes dependent with other X_j's where $j \neq i$.

References

[1] *James, William and Stein, Charles*. [*Estimation with quadratic loss*]. Breakthroughs in statistics (Springer), 443–460, 1992.