Regularization (cont.) and the Bias-Variance Decomposition

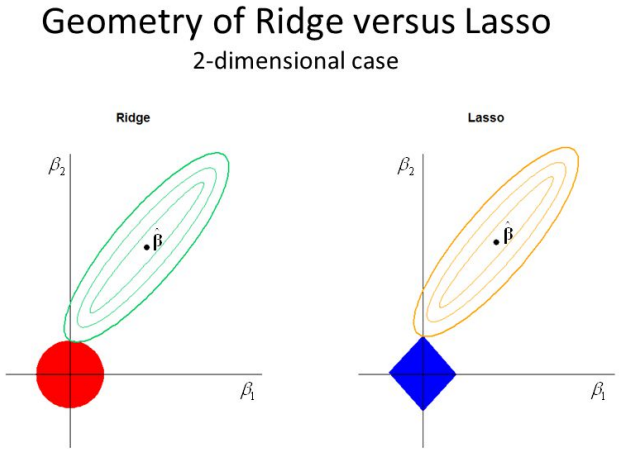
Somani, Neel - Bao, Jason

February 13, 2018

1 Feature Engineering (cont.)

We started today by reviewing feature engineering, by thinking of a variety of features that could be used to price housing for Airbnb.

We demonstrated how we could use LASSO as a means of eliminating features (dimensionality reduction). By looking at the geometry of ridge vs. LASSO, we can see how LASSO promotes feature sparsity (using less features).



Solid areas represent the constraint regions $\beta_1^2 + \beta_2^2 \le t$ & $|\beta_1| + |\beta_2| \le t$

The ellipses represent the contours of the least square error function

2 The Bias-Variance Decomposition: Introduction

We've identified that there's a sweet spot for the number of features, where the testing error is minimized. We don't want too few features, because our model won't be expressive enough, but we don't want too many features either, because our model will overfit our training data and fail to generalize to unseen data points.

We've seen that regularization parameters behave similarly. If we set our regularization parameter too high, then we penalize "complexity" too much, and our model loses expressive power. Fundamentally, the problem is the same.

And so we've reached one of the most important topics in this course: the bias-variance tradeoff.

Bias is the tendency of an estimator to systematically over- or under-estimate a measurement. Formally:

$$Bias_{y^*} = E[y^*] - E[y]$$

where y^* is the estimator, and y is the truth.

We have already encountered *variance* in our introductory statistics courses. To reiterate:

$$Var(X) = E[(X - E[X])^2]$$

We covered the derivation on Wikipedia in class.

3 The Bias-Variance Decomposition: Intuition

Consider the set of all linear combinations of $x_1, ..., x_n$. What geometric figure does it form? It's a plane.

Now consider the least squares predictor over this set:

$$\hat{y} = x * w + b$$

Where is $\hat{y}$ located? It's parallel to the plane of linear combinations.

How do we calculate the least squares predictor? We use our X and Y samples:

$$w^* = X * (X^T * X)^{-1} * X^T * Y$$

The Y samples, though, aren't actually the truth. They likely have some *noise*:

$$y = y^* + z$$

where z is some noise in our measurements. y^*, in the plane of linear combinations that we've drawn, will be above/below the plane.

Now let's recall the equation for a projection:

$$proj_b a = \frac{\langle a, b \rangle}{\langle b, b \rangle} * b$$

Do you see any similarities between the least squares optimum for w^*, and the equation for a projection? They're essentially the same, since $a^T * b$ can be an inner product. In other words, the OLS optimum for w is the projection of Y onto X.

Since the orthogonal projection forms a right triangle, we can observe a relationship between the truth, our predictor, and the error (difference between our predictor and the truth) via the Pythagorean theorem, which is evident in our derivation of the bias-variance tradeoff:

$$\begin{aligned}
\mathbf{E}\left[(y - \hat{f})^2\right] &= \mathbf{E}[y^2 + \hat{f}^2 - 2y\hat{f}] \\
&= \mathbf{E}[y^2] + \mathbf{E}[\hat{f}^2] - \mathbf{E}[2y\hat{f}] \\
&= \mathbf{Var}[y] + \mathbf{E}[y]^2 + \mathbf{Var}[\hat{f}] + \mathbf{E}[\hat{f}]^2 - 2f\,\mathbf{E}[\hat{f}] \\
&= \mathbf{Var}[y] + \mathbf{Var}[\hat{f}] + (f^2 - 2f\,\mathbf{E}[\hat{f}] + \mathbf{E}[\hat{f}]^2) \\
&= \mathbf{Var}[y] + \mathbf{Var}[\hat{f}] + (f - \mathbf{E}[\hat{f}])^2 \\
&= \sigma^2 + \mathbf{Var}[\hat{f}] + \mathbf{Bias}[\hat{f}]^2
\end{aligned}$$

And so we've identified that error comes from three sources: bias, variance, and the irreducible (measurement) error.

We finished the lecture with probability review, including linearity of expectation, independence, and so forth.