Stat 435 Intro to Statistical Machine Learning

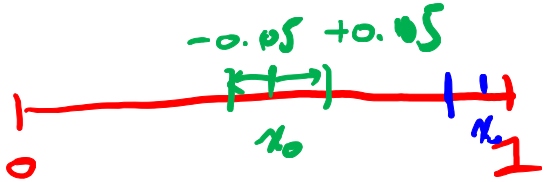
Week 6: Additional exercises

Richard Li

May 3, 2017

Curse of dimensionality

Suppose that we have a set of observations, each with measurements on $p = 1$ feature, X. We assume that X is uniformly (evenly) distributed on $[0, 1]$. Associated with each observation is a response value.
Suppose that we wish to predict a test observations response using only observations that are within 10% of the range of X closest to that test observation.
For instance, in order to predict the response for a test observation with $X = 0.6$, we will use observations in the range $[0.55, 0.65]$. On average, what fraction of the available observations will we use to make the prediction ?



$$\int_0^{0.05} x + 0.05 \, dx + \int_{0.05}^{0.95} 0.1 \, dx + \int_{0.95}^{1} 1 - (x - 0.05) \, dx$$

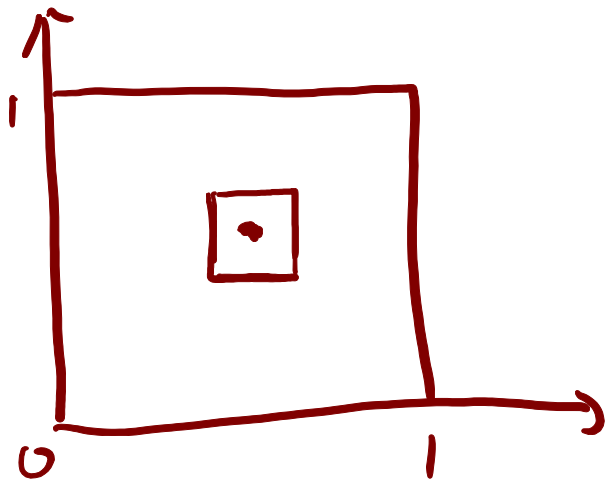
$$= 0.0975$$

Curse of dimensionality

Now suppose that we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2. We assume that (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observations response using only observations that are within 10% of the range of X_1 and within 10% of the range of X_2 closest to that test observation. On average, what fraction of the available observations will we use to make the prediction ?

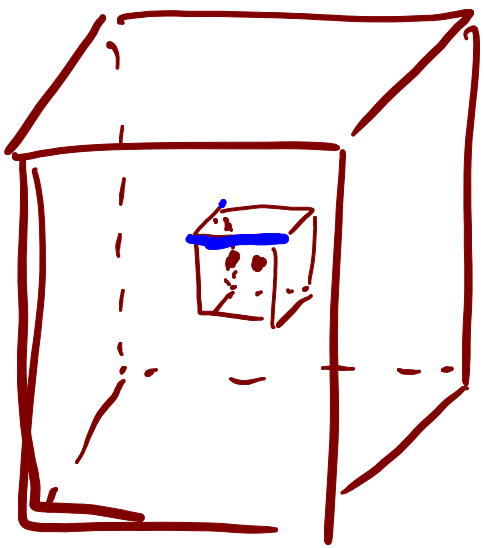
$0.1^2 = 0.01$

$0.0975^2 \approx 0.01$



Curse of dimensionality

What if $p = 100$?



$p = 3$, $0.1^3 = 0.001$

$p = 100$, $0.1^{100} \approx 0$

Drawback of KNN:

There's no "close" neighbour when p is large

Curse of dimensionality

If we want to find a hypercube containing 10% of data. the length of each side should be at least?

$$l^p / 1^p = 0.1$$

when $p = 100$ $\quad l = 0.1^{\frac{1}{p}} \approx 0.9977$

Comparing resampling methods

Consider a very simple model,

$$Y = \beta + \epsilon$$

where Y is a scalar response variable, β is an unknown parameter, and ϵ is a noise term with $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$. Assume that we have n observations with uncorrelated errors. Show that

1. Validation set approach over-estimate the expected test error.
2. LOOCV does not substantially over-estimate the expected test error, provided that n is large.
3. K-fold CV provides an over-estimate of the expected test error that is somewhere between the other two approaches

Comparing resampling methods

Comparing resampling methods

Comparing resampling methods

Comparing resampling methods

Additional question: How about variance of test error?

Shortest review of linear algebra

$tr(ABC) = tr(CAB)$

If you google matrix calculus, this is the picture google shows me:

$\mathbf{y}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$
$\mathbf{Ax}$	$\mathbf{A}^T$
$\mathbf{x}^T\mathbf{A}$	$\mathbf{A}$
$\mathbf{x}^T\mathbf{x}$	$2\mathbf{x}$
$\mathbf{x}^T\mathbf{Ax}$	$\mathbf{Ax} + \mathbf{A}^T\mathbf{x}$

$$\frac{\partial X\beta}{\partial \beta} = X^T$$

$$\frac{\partial \beta^T \Sigma \beta}{\partial \beta} = \Sigma\beta + \Sigma^T\beta$$

(if Σ symmetric)

$$= 2\Sigma\beta$$

(really this is basically all you need to know...)

Projection matrix

The OLS coefficient of the linear regression is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

"hat matrix"

The fitted values are

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY$$

Something interesting about the linear algebra here:

$$HH = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T$$

$$= X(X^T X)^{-1} X^T$$

$$= H$$

Projection matrix

$$\hat{e} = Y - \hat{Y} = Y - HY = (I - H)Y$$

$$\hat{e}^T X = \left[(I-H)Y\right]^T X = Y^T (I-H)^T X$$

$$= Y^T (I - H) X \quad (*)$$

- $IX = X$
- $HX = X(X^T X)^{-1} X^T X = X$

$$\hat{e}^T X = Y^T (X - X) = 0$$

$$\hat{e}^T \hat{Y} = \hat{e}^T X \hat{\beta} = 0 \, \hat{\beta} = 0$$

Ridge regression

Show that the hat matrix associated with ridge regression is not a projection matrix.

$H = HH$

$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y$

$H = X(X^T X + \lambda I)^{-1} X^T$

$HH = X(X^T X + \lambda I)^{-1} X^T \times (X^T X + \lambda I)^{-1} X^T$

$= X(X^T X + \lambda I)^{-1}(X^T X + \lambda I)(X^T X + \lambda I)^{-1} X^T$

$- X(X^T X + \lambda I)^{-1}(\lambda I)(X^T X + \lambda I)^{-1} X^T$

$= X(X^T X + \lambda I)^{-1} X^T - \lambda X(X^T X + \lambda I)^{-2} X^T$

$= H - \lambda X(X^T X + \lambda I)^{-2} X^T \neq H$ unless $\lambda = 0$

Ridge regression

$$\hat{\varepsilon}^T \hat{Y} = (Y - \hat{Y})^T \hat{Y} = (Y - HY)^T HY$$

$$= (Y^T - Y^T H^T) HY$$

$$= Y^T (I - H) HY$$

$$= Y^T (H - HH) Y$$

$$\hat{\varepsilon}^T \hat{Y} = Y^T \left(\lambda X (X^T X + \lambda I)^{-2} X^T \right) Y$$

$$\neq 0 \quad \text{unless} \quad \lambda = 0$$

Regression with correlated predictors

It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting.

Suppose $n = p = 2$, $x_{11} = x_{12}$, $x_{21} = x_{22}$, and both predictors and the response are centered to 0, so we do not have to estimate the intercept.

(a) Write out the ridge regression optimization problem in this setting.

Regression with correlated predictors

Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$

Regression with correlated predictors

Regression with correlated predictors

Argue that in this setting, the lasso coefficient estimates are not unique