Overview of Statistical Learning

Justin Post

September 5, 2014

Statistical Learning

A set of tools, procedure, and theory for modeling and understanding complex datasets.

Encompasses

- Data mining - Finding patterns in datasets
- Inference - Determining which ‘inputs’ (predictors) are associated with the 'output' (response)
- Prediction - Finding the best prediction method for an output based on a set of input variables

Two Major Situations for Learning

- Learner - Model used

- Supervised Learning - Goal is to predict the value of an output (response) based on a number of inputs (predictors or features)

Examples of Supervised Learning

Supervised Learning - Goal is to predict the value of an output (response) based on a number of inputs (predictors or features)

- Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + E_i$$

- Ordinary Least Squares (OLS) estimates given in matrix form by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

where **X** is called the 'design matrix'

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ \vdots & \vdots & \vdots & ... & \vdots \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix}$$

Examples of Supervised Learning

- Goal determines if ‘prediction’ or ‘inference’ is most important

- Possible model fitting procedures (other than OLS):
    - To improve prediction - Ridge Regression, Model Averaging, Neural Networks (Projection Pursuit Regression), ...
    - For better interpretation - LASSO, Elastic Net, Best Subset Regression, ...

Examples of Supervised Learning

Economic survey of Pakistan over the years 1974-75 to 2000-01, yielding 28 observations (Pasha and Akbar Ali Shah, 2004).

Response $=$ # of persons employed (in millions)

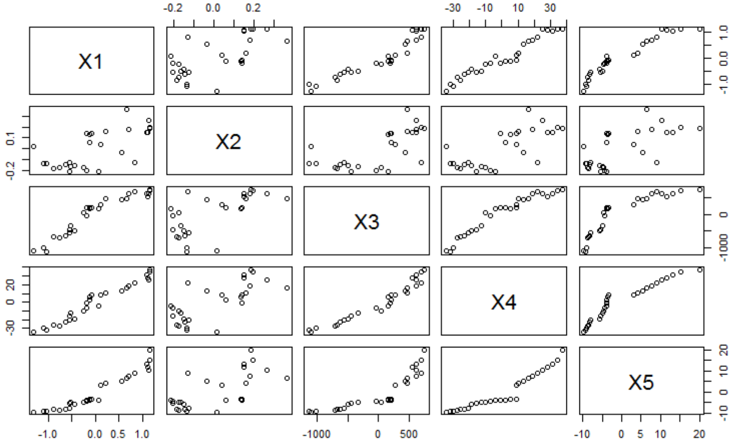
Five predictors:
x_1 $=$land cultivated (in million hectors)
x_2 $=$ inflation rate
x_3 $=$ number of establishments
x_4 $=$ population (in millions)
x_5 $=$ literacy rate

- Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_2 x_{i3} + \beta_2 x_{i4} + \beta_5 x_{i5} + E_i$$

Examples of Supervised Learning

Note: OLS involves $(\mathbf{X}^T\mathbf{X})^{-1}$



Examples of Supervised Learning

Note: OLS involves $(\mathbf{X}^T\mathbf{X})^{-1}$

$$Cor(\mathbf{X}) = \begin{array}{c} \\ \\ \\ \end{array}\left(\begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \\ 1 & & & & \\ 0.664 & 1 & & & \\ 0.943 & 0.659 & 1 & & \\ 0.976 & 0.729 & 0.963 & 1 & \\ 0.956 & 0.681 & 0.867 & 0.951 & 1 \end{array}\right).$$

$(\mathbf{X}^T\mathbf{X})^{-1}$ will be nearly singular! OLS estimates highly unstable.

Ridge Regression can help stabilize estimates and increase prediction accuracy when predictors are correlated! (We'll come back to this.)

Two Major Situations for Learning

- Learner - Model used
- Supervised Learning - Goal is to predict the value of an output (response) based on a number of inputs (predictors or features)
- Unsupervised Learning - No outcome measure. Goal is to describe the associations and patterns among a set of input measures.

Example of Unsupervised Learning

Unsupervised Learning - No outcome measure. Goal is to describe the associations and patterns among a set of input measures.

- Principal Components - Given p inputs $(x_1, x_2, ..., x_p)$, attempt to find linear combinations that best 'represents' the data
  Find
  $$v_1 = a_0x_1 + a_1x_2 + ... + a_px_p = a^T\mathbf{x}$$
  so that $Var(a^T\mathbf{x})$ is maximized.

  Then find
  $$v_2 = b_0x_1 + b_1x_2 + ... + b_px_p = b^T\mathbf{x}$$
  so that $Var(b^T\mathbf{x})$ is maximized subject to v_1 being orthogonal to v_2.

  Repeat until 'enough' of the variation in the $x's$ is described.

Example of Unsupervised Learning

Using Pakistan economic data, we could find the principal components representation of the data.

Since the data are very correlated, perhaps we can find a linear combination (or two) that accounts for most of the variation.

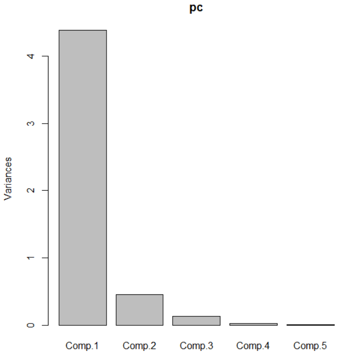
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
SD	2.093	0.672	0.366	0.149	0.100
Prop of Var	0.876	0.090	0.027	0.004	0.002
Cum Prop	0.876	0.967	0.994	0.998	1.000

First PC vector:

$$v_1 = -0.467x_1 - 0.375x_2 - 0.456x_3 - 0.474x_4 - 0.458x_5$$

Example of Unsupervised Learning

Screeplot:

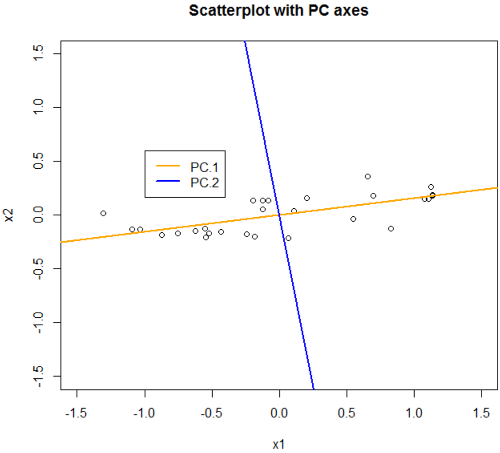


First PC vector:

$$v_1 = -0.467x_1 - 0.375x_2 - 0.456x_3 - 0.474x_4 - 0.458x_5$$

Visualization of P.C.

Consider just x_1 and x_2. Below is a scatterplot with the PC directions plotted.



Two Major Situations for Learning

- Learner - Model used
- Supervised Learning - Goal is to predict the value of an output (response) based on a number of inputs (predictors or features)
- Unsupervised Learning - No outcome measure. Goal is to describe the associations and patterns among a set of input measures.
    - Semi-supervised Learning - Some observations have an output variable, others do not

Supervised Learning

Readings: Most all of both Elements and Introduction cover Supervised Learning

Most problems can be classified as either

- Classification - output variable is categorical

  or

- Regression - output variable is quantitative

Both can be viewed as a task in function approximation.

Selecting a Learner

Consider having all quantitative inputs and output.

Given values of $\mathbf{X}$, we desire a function, $f(\mathbf{X})$, for predicting Y.

Also require a 'Loss function' for determining adequacy of fit, $L(Y, f(\mathbf{X}))$.

Most commonly used Loss function is squared error

$$L(Y, f(\mathbf{X})) = (Y - f(\mathbf{X}))^2$$

Now, criterion for selecting $f()$ is to minimize expected loss.

$$E\left[(Y - f(\mathbf{X}))^2\right]$$

Selecting a Learner

Often a good choice for $f()$ is $E(Y|X = x)$.

This is the solution for Linear Regression:

$$\hat{y} = E(Y|X = x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p$$

A simple solution that works in many situations.

However, when selecting a model there is a constant trade-off between the bias in the model and the variability of the prediction (estimates).

Bias/Variance Trade-off

Linear model, $f(\mathbf{X}) \approx \mathbf{X}^T\beta$, often has low variance but possibly high bias wrt f.

Consider classifying a binary response (0,1) with 2 quantitative pred's.

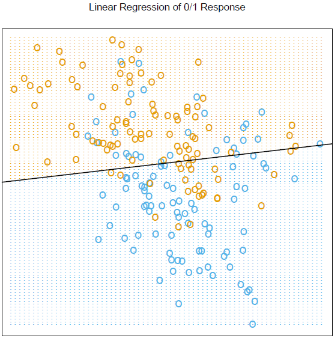


FIGURE 2.1. *A classification example in two dimensions. The classes are coded as a binary variable* (BLUE $= 0$, ORANGE $= 1$)*, and then fit by linear regression. The line is the decision boundary defined by $x^T\hat{\beta} = 0.5$. The orange shaded region denotes that part of input space classified as* ORANGE*, while the blue region is classified as* BLUE.

Bias/Variance Trade-off

Using a local method, such as ‘k-nearest neighbors’ (knn) has low bias but high variance.

- Use k 'closest' (judged by say, Euclidean distance) values to determine classification (0 or 1).
- If knn have proportion of 1's greater than 0.5, assign 1.
- If knn have proportion of 1's less than or equal to 0.5, assign 0.

Bias/Variance Trade-off

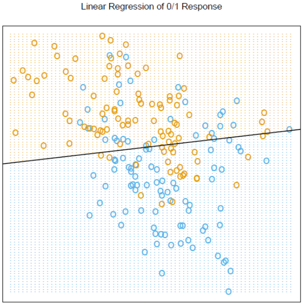


FIGURE 2.1. *A classification example in two dimensions. The classes are coded as a binary variable* (BLUE $= 0$, ORANGE $= 1$)*, and then fit by linear regression. The line is the decision boundary defined by $x^T\hat{\beta} = 0.5$. The orange shaded region denotes that part of input space classified as* ORANGE*, while the blue region is classified as* BLUE.

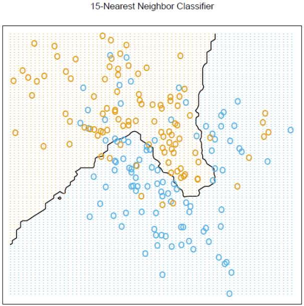


FIGURE 2.2. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (BLUE $= 0$, ORANGE $= 1$) *and then fit by 15-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the 15-nearest neighbors.*

General Methods for choosing $f()$

Parametric - Assume a form for $f()$ leaving us estimation of the resulting parameters

1. Make an assumption about function form. Ex:

$$f(X) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

2. Find parameter estimates by minimizing expected loss. Since we can't actually do this we minimize observed loss. Ex:

$$min_{\beta' s} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2$$

Main drawback, functional form may be incorrect!

General Methods for choosing $f()$

Non-Parametric - No functional form chosen, but put constraint on the ‘wiggly-ness’ of the function.

- Can lead to better ‘fit’ as a wide variety of f’s can be used.
- Problem difficult as f can be complicated
- Need many observations!

Curse of Dimensionality - Any method that attempts to produce locally varying functions in small neighborhoods will run into problems in high dimensions as ‘local’ becomes prohibitively big.

Ridge Regression Example

Pakistan example -

- Assume linear model is true f. MLR model in which errors have mean zero and are uncorrelated with constant variance.
- $\rightarrow$ OLS is 'best' (smallest variance) linear unbiased estimates (by Gauss-Markov Theorem).
- X_1-X_5 highly correlated. Leads to high variance of estimates.
- Perhaps we can get a better model by trading having a little bias, but less variance.
- RR increases bias but decreases variance .

Ridge Regression Example

RR estimates minimize a penalized loss function:

$$min_\beta \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2 + \lambda \sum_{i=1}^{n} \beta_i^2$$

or

$$\hat{\boldsymbol{\beta}}_{RR} = min_{\boldsymbol{\beta}}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

where λ is a 'tuning parameter.'

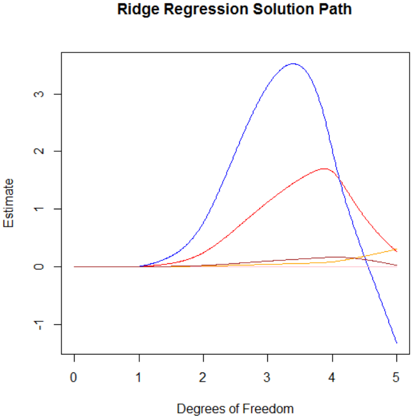
If λ=0, then there is no penalty and you get the usual OLS solution.

If λ is big, then β values that minimize this will be 'shrunk'.

Ridge Regression Example

RR solution has nice matrix form

$$\hat{\beta}_{RR} = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}$$



Principal Components and Ridge Regression

- Principal component with the smallest variance has its coordinates shrunk more.
- This can be seen using the predicted values:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}_{RR} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}$$

Using the SVD of $\mathbf{X}$

$$\mathbf{X}^T\mathbf{X} = (\mathbf{UDV}^T)^T(\mathbf{UDV}^T) = \mathbf{VD}^2\mathbf{V}^T$$

We now have

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{UD}(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{DU}^T\mathbf{y} \\ &= \sum_{j=1}^{p}\mathbf{u}_j\frac{d_j^2}{d_j^2+\lambda}\mathbf{u}_j^T\mathbf{y}\end{aligned}$$

d_j is the j^{th} eigenvalue and $\mathbf{u}_j$ is the j^{th} normalized prin comp of $\mathbf{X}$.

Principal Components Regression

- Principal component regression is an alternative to shrinking all eigenvectors some.
- Here, shrink the smallest eigenvectors to 0 and leave the largest ones untouched.
- That is, simply conduct a regression using the 'most important' eigenvectors as your predictors.

$$Y_i = \beta_0 + \beta_1 v_{i1} + ...\beta_k v_{ik} + E_i$$

where $k \leq p$.

Upcoming Reading Group Topics

- September 19, Brian Gaines - Overview of Regularization Methods
  - Following 2 weeks' topics will be more in depth on Regularization
- October 17, Neal Grantham - Overview of Classification Methods
  - Following 2 weeks' topics will be more in depth on Classification
- November 7, Jami Jackson - Overview of Support Vector Machines
  - Following 2 weeks' topics will be more in depth on SVM

Stat Faculty working in Statistical Learning

- Hua Zhou
- Eric Laber
- Rui Song
- Jessie Jeng
- Yichao Wu
- Howard Bondell
- Dave Dickey
- Wenbin Lu