STAT 444 Statistical Learning

Austin Xia

May 26, 2020

Contents

List of Figures

List of Tables

1 Course Information

1.1 Contact

Instructor: Reza Ramezan
Email: rramezan@uwaterloo.ca

1.2 Grade

Assignments 70%
Quizzes 30%

2 Global modelling methods

2.1 Quick review of linear regression

2.1.1 Simple Linear Regression

$$\mu(x_i) = \mu_{0i} + \mu_{1i}(x_i)$$

regression line is

$$\hat{\mu}(x_i) = \hat{\mu}_{0i} + \hat{\mu}_{1i}(x_i)$$

$y - \mu(x_i)$ and $y - \hat{\mu}(x_i)$ are not the same thing

2.1.2 Multiple Linear Regression

$y = \mu(x) + r$ where $\mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ here comes p value and t score

$$Y|X = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Y is response vector
β is parameter
ϵ is error terms
X is design matrix

2.1.3 Least squares estimation

Definition 2.1.1 (RSS)

$$RSS(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = (y - X\beta)^T (y - X\beta)$$

SSE and RSS are the same thing, both referring to the Residuals sum of squares

Definition 2.1.2 (Ordinary least square estimate)

$$\hat{\beta} = argmin_{\beta}(RSS(\beta)) = argmin\left((y - X\beta)^T(y - X\beta)\right)$$

$$\frac{d}{d\beta}RSS(\beta) = -2X^T(y - X\beta)$$
$$X^T(y - X\beta) = 0$$
$$X^Ty = X^TX\beta$$
$$\beta = (X^TX)^{-1}X^Ty$$

It's normally asuumed that:

- The mean of Y is linear function of X
- Error terms have constant mean 0
- Error terms have constant variance σ^2
- Error terms follow a normal distribution
- Error terms are independent

Need these assumptions to do prediction

Theorem 2.1.1
Assume we can write y in terms of its mean and an error term

$$Y = E(Y|X) + \epsilon = \beta_0 + \sum \beta_i X_i + \epsilon$$

where $\epsilon \sim MVN(\beta, \sigma^2 I_{n*n})$ then

$$\tilde{\beta}_{OLS} \sim MVN(\beta, \sigma^2(X^TX)^{-1})$$

$$(n-p-1)\frac{\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$$

Definition 2.1.3 (The Hat Matrix)

$$\hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^Ty = Hy$$

We have:

- $H^2 = H$
- H(1-H)=1
- tr(H)=p+1

$\hat{r}$ verses ϵ

- $\epsilon = y - X\beta$
- $r = y - \hat{y} = y - X\hat{\beta}$
- r is observed residual, ϵ is model error terms
- We will use $\hat{r}$ as approximate values for ϵ to check if
  - $E(\epsilon_i) = 0$
  - $Var(\epsilon_i) = \sigma^2$
  - $\epsilon_1, ..., \epsilon_n$ are Normally distrubuted
  - $\epsilon_1, ..., \epsilon_n$ are independent

2.1.4 Alternative methods to ordinary least squares regression

If assumptions on error term don't hold true:

- Use transformation to project onto a space where assumption hold true
- Use weighted least squares regression
- Use non-parametric models. e.g. kernel regression, splines, etc.

2.1.5 Influential Points

Based on hat matrix H $\hat{y} = Hy$, we measure influnce by:

$$\frac{d\hat{y}_i}{dy_i} = H_{ii}$$

which is influnence of y_i on the data.
If high leverage points exist in data, one can use robust regression model
Multicolinearity, Curse of Dimensionality $(X^TX)^{-1}$ must exsit, so under multicollinearity or $p \geq n$, LS estimation breaks down.
or if rows are closed to linearly dependent, trace of Hat Matrix get to large, causing prediction having variance too large

2.1.6 Reducible vs. irreducible errors

- prediction is for unobserved data, it needs test set
- inference. We would like to answer these questions in light of sampling variability.

$$MSE = E\left([Y - \hat{Y}]^2|X\right) = \left(f(x) - \hat{f}(x)\right)^2 + Var(\epsilon)$$

the first term is Reducible error, could be 0 if $\hat{f} == f$
the second term is irreducible error. caused by projecting Y into space of X

2.1.7 Bias-Variance tradeoff

$$E\left([Y_0 - \hat{f}(x_0)]^2\right) = E[(f(x_0) - E(\hat{f}(x_0))]^2 + E[(\hat{f}(x_0) - E(\hat{f}(x_0))]^2 + Var(\epsilon)$$

which is

$$Bias(\hat{f}(x_0))^2 + Var(\hat{f}(x_0))^2 + Var(\epsilon)$$

Bias decrease, flexibility increase as complexity increase
Variance increase, stability decrease as complexity increase

2.1.8 Cross-Validation: Basic idea

Definition 2.1.4 (Expected error)
$E(L(Y,\hat{f}(X))$ where $\hat{f}$ is estimate of f, L (loss function) measures distance between Y and $\hat{f}(X)$ eg squared error loss function is defined as

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

Cross Validation forms many test/training sets and measure the prediction error on each test set. The average of these error setimates the expected prediction error

Definition 2.1.5

- Partition data ramdomly to k disjoint and equal-size sets $T_1,...,T_n$ each of size $n_k = n/k$
- For $i=1,...,k$ construct training sets $T_{-i} = T_1 ... \cup T_{i-1} \cup T_{i+1} \cup ... T_n$ calculate
$$sMSE_i = \frac{1}{n_k}\sum_{j\in T_i}\left(y_j - \hat{f}^{-i}(x_j)\right)^2$$
where $f^{-i}(x_j)$ is estimate/fitted value of y_i based on a model fitted to T_{-i}
- the overall k-fild cross-validation error is
$$CV_k = \frac{1}{k}\sum_{i=1}^{k} sMSE_i$$

Note: loss function can be replaced
it can be shown that

$$CV_k = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{f}^{-\omega(j)}(x_j)\right)^2$$

this formula is understood base on test set
$\omega : 1...n \to 1...k$ is an indexing function that inducates the partition to which observation j is allocated by the ramdomization
choice of k is bias-variance trade-off. Large k results in similar training sets, high variance, smaller bias, more flexibility
in practice, k is set to 5/10

2.1.9 Leave-One-Out CV: LOOCV

- k=n in k-fold cross validation. i.e. training sets of size n-1 and test sets of size 1
- it has largest k, so very little bias and large variance low prediction power.
- it's computationally efficient. For least squres regression, it can be shown that
$$CV_n = \frac{1}{n}\sum_{i=1}^{n} sMSE_i = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\hat{y}_i}{1-H_{ii}}\right)^2$$
The result above is true for most linear smoothers $\hat{Y} = SY$ under squared error loss, in which case H_{ii} will be repalced by S_{ii}

2.1.10 Generalized Cross-Validation:GCV

A convenient approcimation to LOOCV for linear fitting under squared-error loss. The GCV approximation of the prediction error is

$$GCV_n = \frac{1}{n} \sum \left(\frac{y_i - \hat{f}(x_i)}{1 - trace(S)/n} \right)$$

trace(S) is effective number of parameters in the model (degree of freedom). so number of parameter increase then GCV decrease prediction power

2.1.11 Best-subset selection

The subset selection is the process of selecting q ¡ p explanatory variates in modelling the function f
Problem with LS estimate $X\hat{\beta} = HY$

- prediction accurary, LS have low bias, but if not $n \gg n$ they have larger variance.
- multicollinearity
- if $p > n$ then LS estimate are not unique
- Model interpretability: including irrelevant variables leads to unneccessary complexity

Best-subset Selection:

- a total of $2^n = \binom{p}{0} + ... + \binom{p}{p}$ models are to be fitted
- An efficient algorithm makes this feasible for p as large as 30-40

Steps:

- M_0 denote the null model, no predictor, the model simply predicts sample mean
- For $k = 1, ..., p$: Fit all $\binom{p}{k}$ models that contain exactly k predictors
  pick the best among these models. call it M_k
- select the best model among them using cross-vadiated prediction error

2.1.12 Forward/Backward-stepwise selection

Forward stepwise selection The algorithm:

- Let M_0 denote the null model, containing no prectors, predicts sample mean
- For k = 0, ..., p-1
  - consider all p-k models that augument the predictors in M_k with k additional predicot
  - choose the best among these p-k models, call it M_{k+1}.
- select the best model

Backward stepwise selection the algorithm: every step there is one less predictor
Hybird Approach:

- both forward and backward
- vriables are added to model, after each new variable, the method may also remove variables

- attempts to mimic subset selection while retaining the computational advantages of forward and backword stepwise selction

Stepwise methods are more constranit than best subset selection, hence they have lower vairance but perhaps more bias
Shrinkage/regularization methods are computationally efficient.

- LS: $min_{\beta}(RSS)$
- Shrinkage"$min_{\beta}(RSS) + \lambda * Pen(\beta)$
  Penalizing

2.1.13 Weighted Least Squares

if we want to give more(or less) weights to different observation if there is Outliers and Heteroscedasticity
Heteroscedasticity (constant variance): $y = X\beta + \epsilon$
$\epsilon_1, ..., \epsilon_n \backsim^{iid} N(0, \sigma^2)$

$$WRSS(\beta) = \sum_{i=1}^{n} w_i \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

$$\hat{\beta_{WLS}} = argmin_{\beta}(WRSS(\beta)) = (X^T W X)^{-1} X^T W y$$

derivation, see slides

2.1.14 choice of W

In a word it's inversly proportional to variance of x_i

- has to do with variance of y (we have to know the variance(for some measuring device i.e.))
- if repeated measurements of Y for each X is available, then $Var(Y_i|X_i) = Var(\epsilon_i)$ can be estimated
- if we can assume distribution. The variance of proprtions we find should be inversly proportional to the sample size
  so a natrual choice of weights is proportional to the sample size

  Choice of W:

  - assumption: Y_i:average of n_i repeated measurements.
    $w_i = kn_i$

2.2 Regularized regression models

2.3 Robust regression and breakdown

2.4 Non-parametric regression

3 Locally adaptive methods(smooth functions)

3.1 Local linear regression

3.2 Smoothing splines

3.3 Kernel method

3.4 Density/intensity function estimate

4 Preditive accuracy

4.1 Roles of training and test data, cross-validation, etc

4.2 Bias/Variance trade-off and parameter choice

5 Locally adaptive methods(tree-based methods)

5.1 regression trees