Mid-term Review

DSA 6000: Data Science and Analytics, Fall 2019

Wayne State University

Chapter 1

- Statistical Concepts:
  - Random Variable, Expected Value, Variance, Probability Distribution
- Textbook notation and simple matrix algebra
  - Matrix-vector multiplication in R
- Machine learning overview (slides for lecture 1)

Chapter 2

- Concept of $Y = f(X) + \epsilon$ and concept of irreducible error
- Distinction between model and estimate, e.g. f and $\hat{f}$
- Be able to tell apart inference v.s. prediction, regression v.s. classification, supervised v.s. unsupervised learning instances, and parametric v.s. non-parametric methods
- Be able to measure the quality of fit using mean squared error and error rare
- Understand the concepts of overfitting, bias-variance tradeoff
- Understand the KNN classifier, can do calculation on Euclidean distance
- Understand the curse of dimensionality

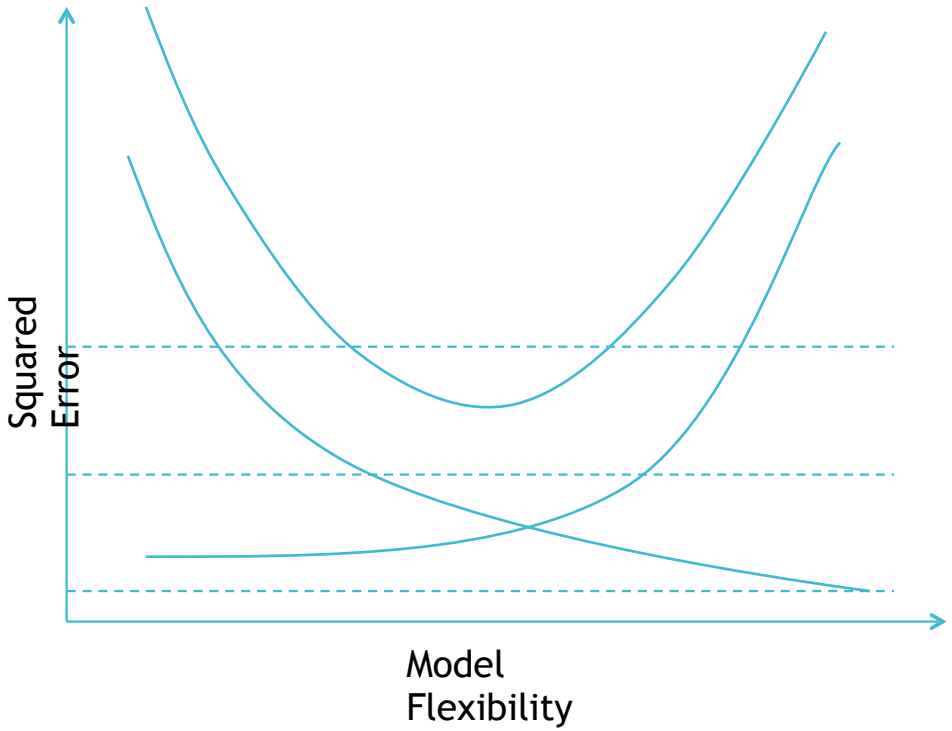
Chapter 3

- Linear regression model assumptions
- Understand the concept of the least squares method
- Understand the partitions of sums of squares, R-square (3.17), RSS (3.16), RSE (3.25)
- Be able to fit a multiple linear regression using lm() and interpret its summary() output
- Be able to make predictions and interpret confidence interval and prediction interval
- Coding of categorical variables, 3.3.1
- Understand the interaction effect of two predictors, 3.3.2
- Linear model diagnostics: residual plot, heteroscedasticity, outlier, high-leverage point, collinearity, VIF, danger of extrapolation

Chapter 4

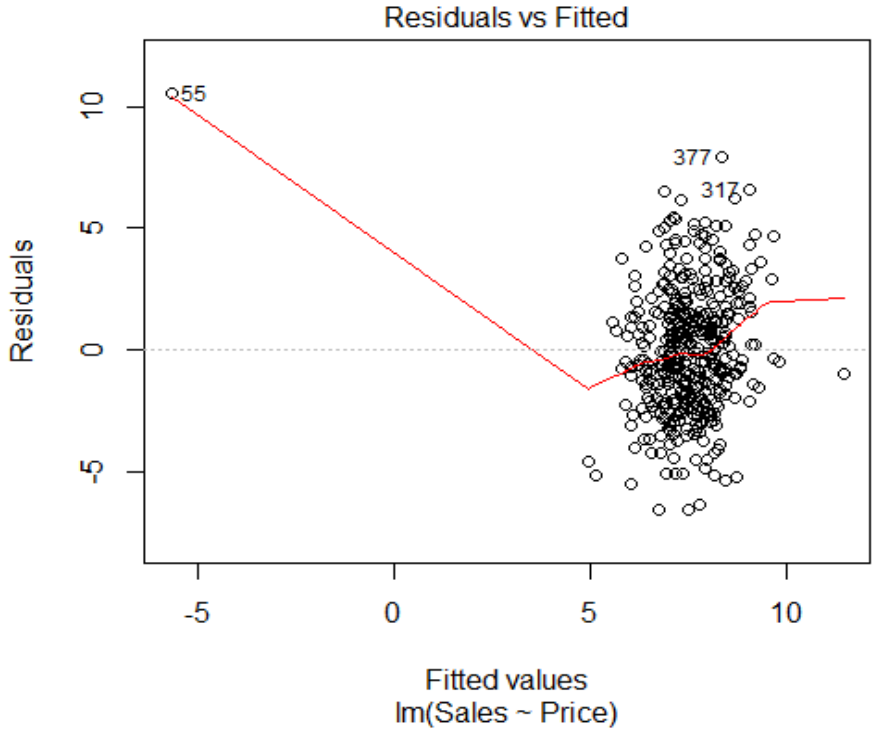
- Logistic regression, formulas, odds, log of odds
- Be able to make predictions using logistic regression
- Interpret coefficients in terms of log of odds, odds and odds ratio
- Understand the concept and procedure of linear discriminant analysis (LDA)
- Application of Bayes theorem and Naïve Bayes classifier
- Assess the classification performance: confusion matrix, accuracy, TPR, FPR, ROC curve, AUC

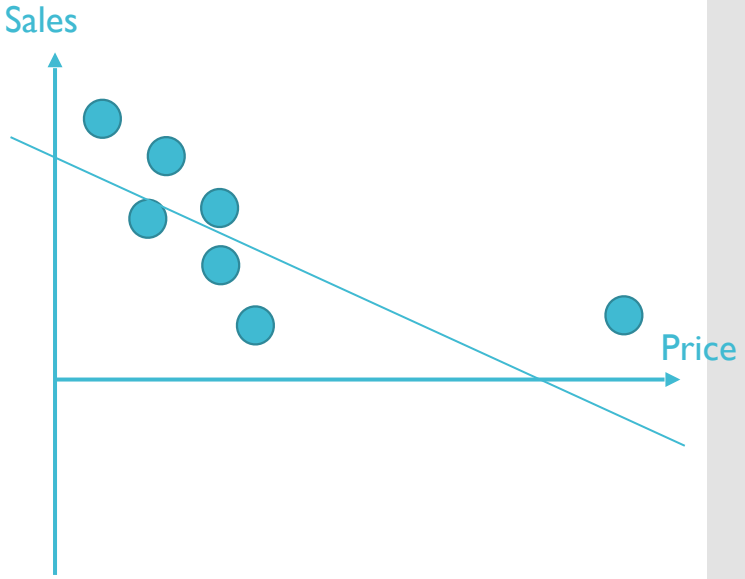
- Label out Bias^2, Variance, Irreducible Error and Test MSE



```
Call:
lm(formula = mpg ~ horsepower + weight + year, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7911 -2.3220 -0.1753  2.0595 14.3527

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.372e+01  4.182e+00  -3.281  0.00113 **
horsepower  -5.000e-03  9.439e-03  -0.530  0.59663
weight      -6.448e-03  4.089e-04 -15.768  < 2e-16 ***
year         7.487e-01  5.212e-02  14.365  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.43 on 388 degrees of freedom
Multiple R-squared:  0.8083,    Adjusted R-squared:  0.8068
F-statistic: 545.4 on 3 and 388 DF,  p-value: < 2.2e-16
```

- What is the mean of the residuals and why?
- Calculate the VIF of weight.
- State the null hypothesis associated with the t-test on horsepower, and make conclusions of the test result.





I built a *simple* linear regression for $Sales = \beta_0 + \beta_1 \cdot Price$ on a training set and generated the residual plot. (Residual = Actual – Fitted)

If the 55th observation is removed from the training data and the model is refit, the slope of the regression line will _____ (Use **increase** or **decrease** to fill in the blank).

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

TABLE 4.3. *For the* Default *data, estimated coefficients of the logistic regression model that predicts the probability of* default *using* balance, income, *and student status. Student status is encoded as a dummy variable* student[Yes], *with a value of* 1 *for a student and a value of* 0 *for a non-student. In fitting this model,* income *was measured in thousands of dollars.*

- What does the coefficient 0.0057 for balance mean?

At a fixed income and student status (regardless what value they take), one dollar increase in balance

1. Corresponds to 0.0057 unit of increase in *log of odds of default* (or logit of default).
2. Corresponds to exp(0.0057) = 1.00572 times change in odds of default, or a 0.572% increase in odds.

Both statements are OK. Note: I am using "increase" to describe the difference, not an action. There is not a causal relation between balance and default. A better way: use person A and person B for a comparative statement.

- You built a logistic regression model to predict loan applicant's risk of default based on information on the application form. When your client (say, a bank) evaluated the model on an internal test data set, they found that while the true positive rate (rate of good applicants being accepted) is quite high, the false positive rate (rate of bad applicants being falsely accepted) is also very high.
- Based on this, your client concludes that your model is crappy. Is that conclusion solid? As a data scientist, where would you start to look into this issue?

- You are a data scientist of a large car dealer chain and have access to the customer database. Your internal business partner tells you to **do a clustering analysis to group the customers into a high-value cohort and a low-value cohort** so he can plan his marketing strategies.
- What is the problem with this request?
- Ask him two follow-up questions toward understand/ addressing his objective.

True or False

- Adding a new independent variable to a multi-linear regression model will always increase the explained variance in the response variable in the training data.
- Correlation always implies causation.
- A more complex model is more prone to overfitting than a simpler model.
- If two variables are correlated, then they have a linear relationship.
- In the KNN classifier, a larger value for K will in general produce a higher variance on the test data.

- The goal of the regression model is to achieve the R-squared value _____

A. Closer to 0
B. Closer to 1
C. More than 1
D. Less than 1

- Which of the following tests can be used to determine whether a linear association exists between the dependent and independent variable in a simple linear regression model?
    A. T-test
    B. ANOVA F-test
    C. Both of the above
    D. None of the above

- Write down a command in R that will produce (and print to the console) the sequence of odd numbers from 1 to 100, i.e.:
  - $1, 3, 5, \ldots, 99$