Chapter 2: Stats Basics

Name	Content
TYPE	cheat sheet
BOOK	An Introduction to Statistical Learning
AUTHORS	Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani
PUBLISHER	Springer

2.1 Stats Basics

This cheat sheet sets up what you need to know to understand basic statistics. It's the *minimum* information you need to use as a reference. Think of it like the 3x5 card you were allowed to bring into tests. In the Notes file of this repo there are more in-depth notes.

Basic relationship between X and Y

$Y = f(X) + \epsilon$
where X is the dependent variable (predictor) and Y is the independent variable (response)

Prediction & Inference

All of statistics is concerned with estimating f and we can do it in two main ways:

1. **Prediction** we know X but not Y.
   the function is a black box - we don't care how it works as long as it's accurate.
   there will always be error.
   a. *reducible error* = error created because we didn't account for a predictor that is related to the response. Better models will help reduce this.
   b. *irreducible error* = the world is messy, statistical models can't capture real life. Nothing can get rid of this error. This gives us an upper bound of the accuracy of our prediction. Almost always unknown in practice.
2. **Inference** We want to know how changes in X affect Y. We do not necessarily want to make predictions for Y.

the function can't be a black box, we have to know exactly how X is impacting Y.
linear models are used for inference.

Statistical Methods

There are two statistical methods we can use:

1. **Parametric** methods make some assumptions about the form f takes.
   may not accurately reflect reality, because it is restricted by the assumptions.
2. **Non-Parametric** do not make explicit assumptions about the from f takes.
   will fit the data better, but also needs more observations to get an accurate representation of f.

Supervised, Unsupervised, & Semi-Supervised Learning

1. **Supervised Learning** several predictor measurements for a given y.
2. **Unsupervised Learning** has no associated y for the predictor measurements.
3. **Semi-Supervised Learning** has observations for the predictors, but the measurements for the response are less available.

Regression vs. Classification Problem

Most statistical methods are based on whether the response, Y, is qualitative or quantitative.

1. **Regression Problems** involve quantitative (numerical) data.
2. **Classification Problems** involve qualitative (categorical) data.

2.2 Assessing Model Accuracy

There is no best method. The data you have determines what method you use.

Quality of Fit

Quality of Fit is the extent to which the predicted response value for an observation is close to the actual response value. (i.e. How close is our predicted y is to the real y for a given x?)

The most common measure is the mean squared error, MSE.

- we want a small test MSE
  - This means we use training data to train the statistical method
  - Then use data *not in the training set* to determine the test MSE.
  - **Cross-Validation:** estimating the test MSE using the training data.

Overfitting is a model that has a small training MSE but a large test $ MSE.

The Bias-Variance Trade Off

The test MSE is the sum of:

1. The variance of $\hat{f}(x_0)$
2. The squared bias $\hat{f}(x_0)$
3. The variance of ϵ

The test MSE will never be lower than the irreducible error $Var(\epsilon)$

1. **Variance** the amount that our predicted f would change if we used different training data.
2. **Bias** the error that's introduced by trying to use a statistical model to approximate reality.

We want low variance and low bias, but often have to make decisions about which one to prioritize.

The Classification Setting

To quantify the accuracy of f when the response Y is qualitative, we use the **error rate**.

- This is a number of how many observations were misclassified.
- We are still interested in the *test* error rate.
- If the test error rate is small, it means we have a good classifier.

The Bayes Classifier

The test error rate is reduced by a simple classifier that assigns each observation to the most likely class, give its predictor.
For example, the probability that a district(observation) has *fraud* (class) given *high turn out* (predictor).

The Bayes Classifier says there are only two possible response values (e.g. fraud, no fraud)

Bayes Decision Boundary where the points have a probability of exactly 50%.
Bayes Error Rate lowest possible test error rate (analogous to the test error rate)

The Bayes Classifier is an unattainable gold standard against which we judge other models.

Bayes Equation

$$P(A|B) = \frac{P(B|A)P(A)}{P|B}$$

A, B = Events
$P(A|B)$ Probability of A given B is true

$P(B|A)$ Probability of B given A is true
$P(A), P(B)$ Independent probabilities of A and B

Bayes Equation Example

- the chances of me hugging a polar bear is rare (1%)
- but giving hugs is fairly common (10%)
- and going to Canada increases my chance of seeing a polar bear (90%)

$$P(HUG|CANADA) = \frac{P(CANADA|HUG)P(HUG)}{P(CANADA)}$$

$$P(HUG|CANADA) = \frac{90\% \times 1\%}{10\%} = 9\%$$

So the probability of me hugging a polar bear if I go to Canada: 9%

K-Nearest Neighbors

Estimates the conditional probabilty $(Y|X)$ then classifies an observation to the class with the highest *estimated* probability.

1. The researcher chooses a value for K (usually $k = 5$ or $k = 10$).
2. Using training data, KNN identifies the K points closes to the test observation using the Bayes rule.

Low K: More flexible, low bias, high variance
High K: Less flexible, high bias, low variance.