Statistics 101C Discussion Week 3

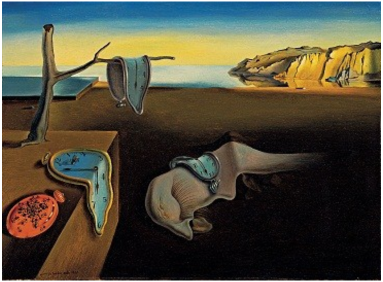
Brice Randolph

University of California, Los Angeles

brandolph@ucla.edu

April 12, 2016

Agenda



What will we be doing in discussion section today?

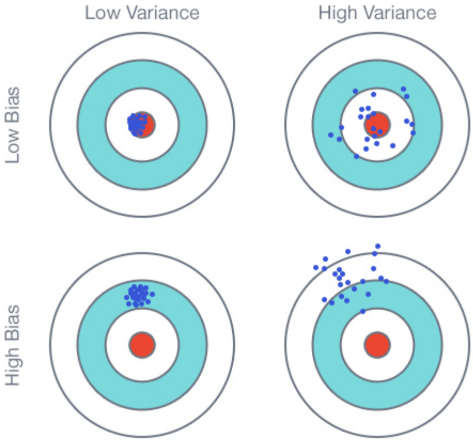
- K nearest Neighbors
- Logistic Regression
- Linear Discriminant Analysis
- Python code + graphics
- Briefly discuss Cython / Rcpp
- Practice coding interview question

The Dude Recommends

- Knitr workflow
- Have you checked stack overflow yet?
- Google python style guide

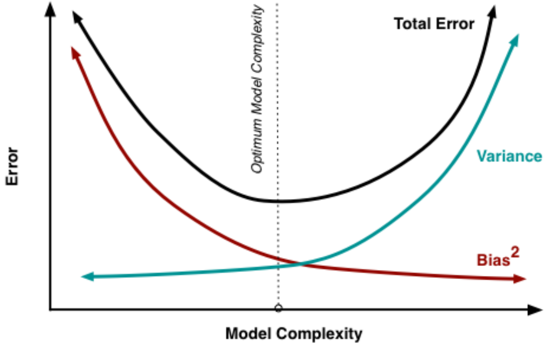
Bias / Variance

As flexibility increases, bias tends to decrease more rapidly than variance increases.

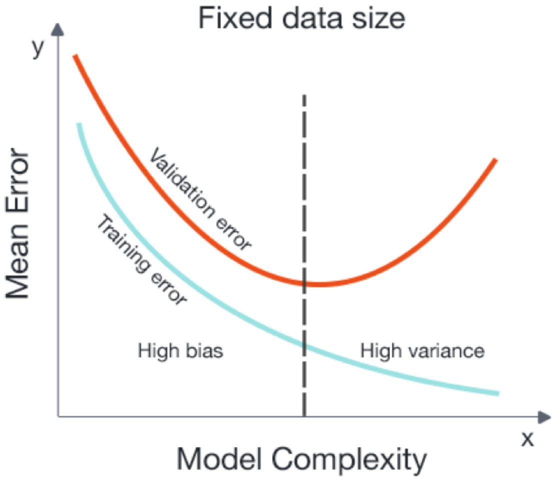


Bias / Variance

As flexibility increases, bias tends to decrease more rapidly than variance increases.



Bias / Variance



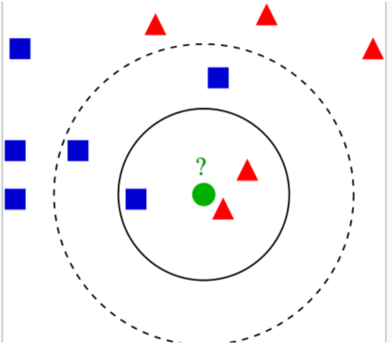
Going beyond linear models

Some things to keep in mind when you see a new model in class:

1. How do I fit this model in R?
2. What are the parameters (hyper-parameters) of the model?
3. Supervised or Unsupervised?
4. Can I explain how this model works to my colleague? What about my parents?

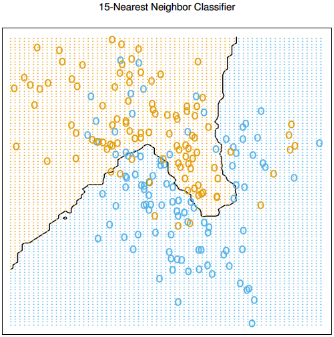
K-th Nearest Neighbor

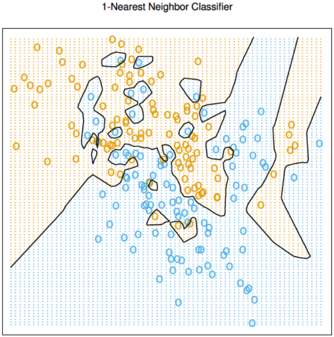
In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.



K-th Nearest Neighbor

1. What is K?
2. How do you initialize the centers?
3. Distance metric?
4. More flexible? *k=1 or k=3
5. As K increases, flexibility is going down

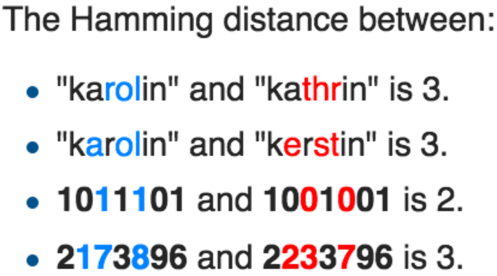




(From **ESL**, chap.2)

Distance Metrics

In text classification, one might be interested in using a different distance metric. The hamming metric is one such choice.



The Hamming distance between:

- "karolin" and "kathrin" is 3.
- "karolin" and "kerstin" is 3.
- **1011101** and **1001001** is 2.
- **2173896** and **2233796** is 3.

library(class)

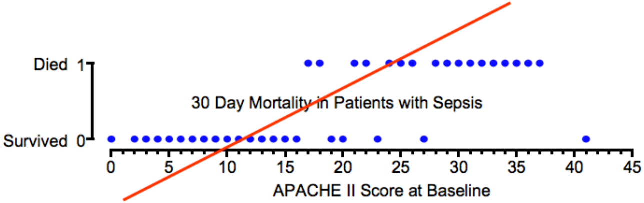
1. knn()
2. training data frame with only predictors
3. testing data frame with only predictors
4. vector for training categorical responses

Logistic Regression

1. What is it used for?
2. want $P(Y == 1)$
3. glm(Y Diagonal, data = banknote, family = binomial)
4. family = gaussian (what would this do?)

Predicting death from baseline APACHE II score in ICU patients

1. APACHE II score can be calculated on all patients newly admitted to the intensive care unit. While it is not mandatory and will not help with patient management, it is a useful tool for risk stratification and to compare the care received by patients with similar risk characteristics in different units.
2. http://www.mdcalc.com/apache-ii-score/



Logistic Regression

1. What do intercepts tell us?
2. $log(\frac{\pi_x}{1-\pi_x}) = \beta_0 + \beta_1 x$
3. predict() gives predicted log odds
4. predict(type = "response")
5. what do negative log odds mean? predicted to be good in counterfeit bill example

Logistic Regression

1. Think of 3 examples where you might use logistic regression for a prediction.
2. Also thing of a couple good predictors in each of those examples.

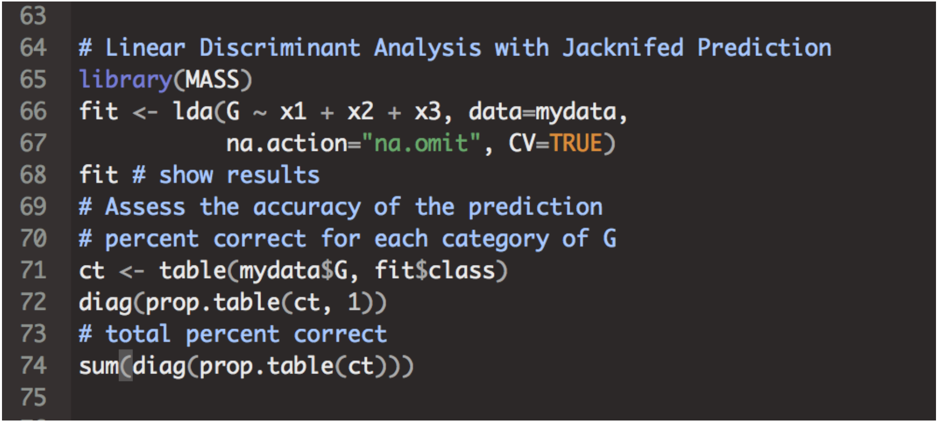
Generative vs Discriminative classification

What's the difference?

1. Generative classifiers learn a model of a joint probability $p(x, y)$, of the inputs x and the label y, and make their predictions using Bayes rule to calculate $p(y \mid x)$, and then picking the most likely label y.
2. Discriminative classifiers model the posterior $p(y \mid x)$ directly, or learn a direct map from inputs x to the class labels.

Linear Discriminate Analysis

What does this code do?



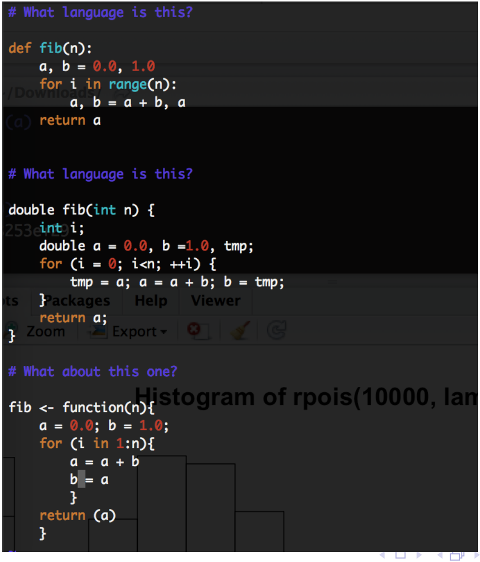
```
# Linear Discriminant Analysis with Jacknifed Prediction
library(MASS)
fit <- lda(G ~ x1 + x2 + x3, data=mydata,
           na.action="na.omit", CV=TRUE)
fit # show results
# Assess the accuracy of the prediction
# percent correct for each category of G
ct <- table(mydata$G, fit$class)
diag(prop.table(ct, 1))
# total percent correct
sum(diag(prop.table(ct)))
```

Extreme Assumptions

1. Normality of predictors?
2. Equal variances of classes?
3. Equal proportions of classes in the population
4. LDA comes in as $x < \frac{\mu_1 + \mu_2}{2}$
5. How do we know the means?

Name that language!

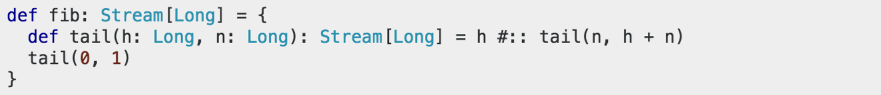
First, what is this function doing?



```
# What language is this?

def fib(n):
    a, b = 0.0, 1.0
    for i in range(n):
        a, b = a + b, a
    return a


# What language is this?

double fib(int n) {
    int i;
    double a = 0.0, b =1.0, tmp;
    for (i = 0; i<n; ++i) {
        tmp = a; a = a + b; b = tmp;
    }
    return a;
}

# What about this one?

fib <- function(n){
    a = 0.0; b = 1.0;
    for (i in 1:n){
        a = a + b
        b = a
        }
    return (a)
    }
```

Name that language!

What does this one look like?



```
def fib: Stream[Long] = {
  def tail(h: Long, n: Long): Stream[Long] = h #:: tail(n, h + n)
  tail(0, 1)
}
```

Practice Job interview Question

Biggest strength?

How would you bring value to the company?

What brought you into statistics/data science?

How do you convince your supervisor that he is making an incorrect decision?

Practice Job interview Question

Use R or Python:

Given a length n list of integers A, write a function that takes this list as an argument, and returns another list where each element is the integer j_i where $1 \leq j_i \leq A_i$ and j_i is the number of integers j that are coprime to A_i.
First think of helper functions that you could use.