Classification, Including Naive Bayes

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

Outline

Announcements

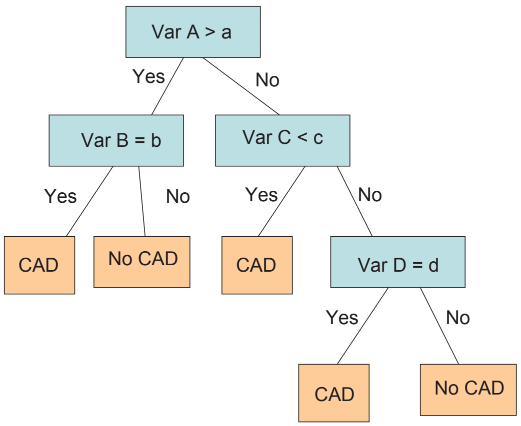
- Before lab next week, register at www.dataminingbook.com so that you can get the data sets
- Review conditional probability, independence, joint probability, Bayes' rule! Will be used heavily in this unit.
- Reading this week: SPB pp. 50-58 & Chap. 8. "Intro to R": Sec. 5.2, 5.7, 7.1, 9.2, 10.0, 10.1, 10.3.
- Questions?

Heart Disease Detection

- The goal is to learn a good classification rule to predict the presence / absence of CAD from the 13 predictors in the data set:
  - age
  - sex
  - chest pain type
  - blood pressure
  - Number of vessels showing calcium on fluoroscopy
  - exercise thallium scintigraphic defects (fixed, reversible, none)
  - electrocardiogram results
  - exercise-induced angina (presence / absence)
  - etc.

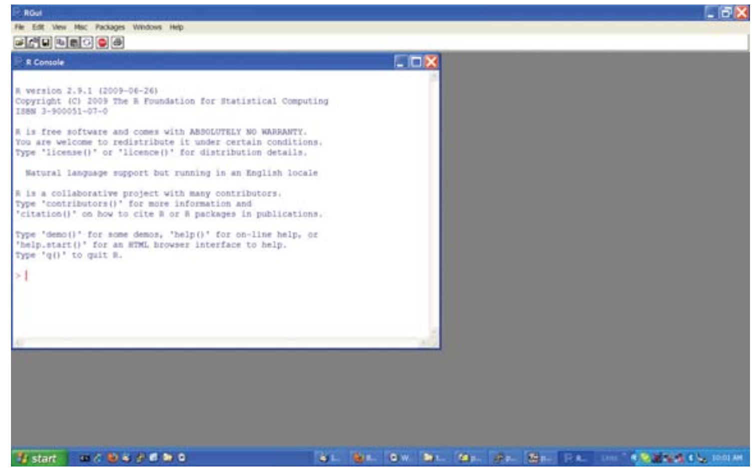
Heart Disease Detection

Let's use a classification tree:



Heart Disease Detection

First open R and read in the Cleveland data:



Heart Disease Detection

- For each categorical predictor, look at the joint counts table for the predictor with the outcome (CAD):

  > table(heart$sex, heart$cad)

	N	Y
Fem	67	20
Mal	83	100

- CLICKER: Based on these data, does gender appear to be associated with heart disease?

  A. Yes
  B. No

Heart Disease Detection

- Obtain the marginal counts for gender:

  > table(heart$cad)

Fem	Mal
87	183

- Use these to obtain the conditional frequencies of CAD given gender:

  > <divide the joint table by the marginal table>

	N	Y
Fem	0.77	0.23
Mal	0.45	0.55

- Clearly the frequency of CAD is very different between men and women (variables are highly dependent)

Heart Disease Detection

- Alternatively, one can look at the conditional frequencies of gender given CAD / no CAD:

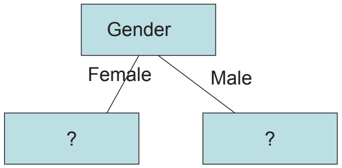
> <code not shown>

	Fem	Mal
N	0.45	0.55
Y	0.17	0.83

- Also shows that the variables are highly dependent (the rows are very different)
- If these variables were independent, what would this conditional frequency table probably look like?

Heart Disease Detection

- Since gender is strongly associated with heart disease in the data, and since I know that physicians treat heart disease differently for men vs. women, I chose the first branch of the classification tree to be gender:



Heart Disease Detection

- In order to decide what variable to branch on for the males, and for the females, split the data by gender:

```
> females = heart[ heart$sex == "Fem", ]
> males = heart[ heart$sex == "Mal", ]
```

- heart$sex selects the variable “sex” in the data set “heart”
- As we will learn in an R tutorial next week, the comma indicates that we are selecting a subset of the rows of the data set (those having a particular value for heart$sex), and selecting all columns of the data.

Heart Disease Detection

- Looking just at the males, find the conditional frequencies of every other predictor variable, given CAD / no CAD.
- Here they are for the predictor, “blood sugar $>$ 120 mg / dl”:

> <code not shown>

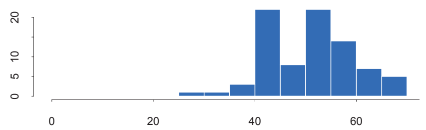
	N	Y
CAD = N	0.80	0.20
CAD = Y	0.88	0.12

- Is this predictor highly dependent with CAD status for males?

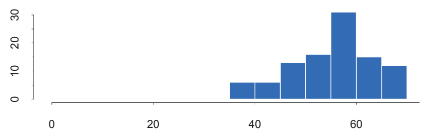
  A. Yes
  B. No

Heart Disease Detection

- For continuous predictor variables, consider the conditional distribution, given CAD / no CAD:



Age of no−CAD males



Age of CAD males

Heart Disease Detection

- Are age and CAD status highly dependent for males?
- Could we use the age variable to distinguish effectively between CAD / no CAD males? I.e. is there a particular age above which almost all males have CAD, and below which almost no males have CAD?

  A. Yes

  B. No

Heart Disease Detection

- Here is the conditional frequency of the predictor “exercise induced angina” for males, given CAD / no CAD:

	N	Y
CAD = N	0.82	0.18
CAD = Y	0.44	0.56

- Is this predictor strongly associated with CAD / no CAD for males?
- Here is the conditional frequency of the predictor “Number of vessels containing calcium” for males, given CAD / no CAD:

	0	1	2	3
CAD = N	0.83	0.11	.02	.04
CAD = Y	0.32	0.37	.19	.12

- Is this predictor strongly associated with CAD / no CAD for males?

Heart Disease Detection

- One could split the males on either “exercise induced angina” or “Number of vessels containing calcium”
- Before I did this, I looked at the joint distribution of the two predictors for men with CAD:

	0	1	2	3
0	0.16	0.15	0.08	0.05
1	0.16	0.22	0.11	0.07

- and the joint distribution of the two predictors for men without CAD:

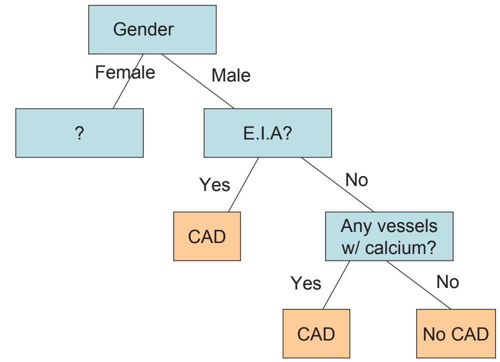
	0	1	2	3
0	0.69	0.07	0.02	0.04
1	0.14	0.04	0	0

Heart Disease Detection

- The combination of no exercise induced angina and no vessels containing calcium was very common among the no-CAD men but uncommon among the CAD men.
- Let's classify the men who have no exercise induced angina and no vessels containing calcium as non-CAD, and all other men as CAD

Heart Disease Detection

- That leads to the tree:

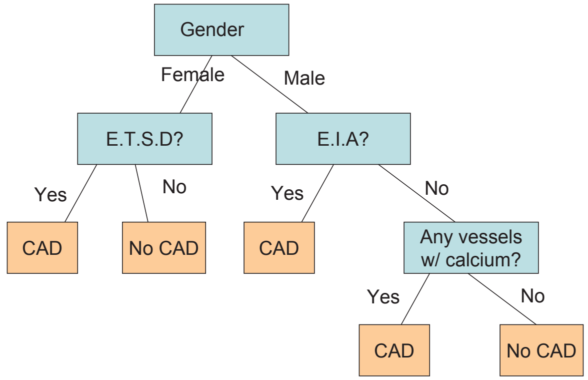


Is there an equivalent tree?

A. Yes
B. No

Heart Disease Detection

- For women, CAD is strongly associated with the predictor, "exercise thallium scintigraphic defects"
- So our final tree is:



Heart Disease Detection

- The tree has 80% classification accuracy on the training data
- Ideally we would evaluate accuracy on the test sets, but the test sets have missing data, so we cannot apply our classification tree
- Heuristic classification methods like this one have trouble with missing data, but many statistical classification methods can handle missing data (we will learn one next week)

Accidents data

- U.S. Department of Transportation data on automobile accidents.
- Whether an injury occurred
- Factors that influence the chance of an injury
  - time of day
  - alcohol involved?
  - speed limit

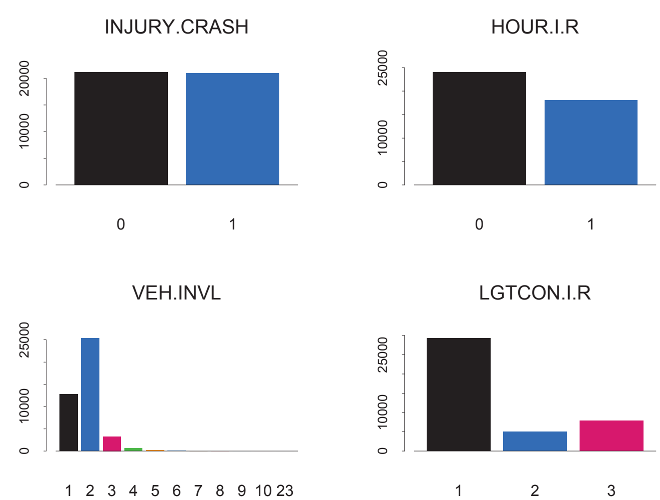
Accidents data

- An emergency response center wants to assign priority levels to accidents based on the chance of an injury, given the immediately available information such as time of day
- This is a classification problem (predict injury / no injury based on predictor vars)
- Is this supervised or unsupervised learning?

The Accidents Data

	A	B	C	D	E	F
1	HOUR_I_R	ALCHL_I	ALIGN_I	STRATUM_	WRK_ZON	WKDY
2	0	2	2	1	0	
3	1	2	1	0	0	
4	1	2	1	0	0	
5	1	2	1	1	0	
6	1	1	1	0	0	
7	1	2	1	1	0	
8	1	2	1	0	0	
9	1	2	1	1	0	
10	1	2	1	1	0	
11	0	2	1	0	0	
12	1	2	1	0	0	
13	1	2	1	1	0	
14	1	2	1	1	0	
15	1	2	2	0	0	
16	1	2	2	1	0	

Accidents data



Naive Bayes: A Classification Method

Probabilistic Approach:

- Want to estimate the conditional distribution of the (categorical/discrete) outcome variable Y given the predictor variables $\{X_k : k = 1, \ldots, K\}$

$$\Pr(Y|X_1, \ldots, X_K)$$

- E.g., the probability of injury given time of day, alcohol involvement, etc.
- That can be used to estimate whether or not an injury occurred at an accident site.

Naive Bayes

- If $\{X_k : k = 1, \ldots, K\}$ and Y are both discrete random variables, one could estimate $\Pr(Y = y|X_1 = x_1, \ldots, X_K = x_K)$ for any combination of y and $x_1, \ldots, x_K$ using the conditional frequencies from the training data
- Why is this NOT a good idea?

Naive Bayes

Alternative:

- Model the joint distribution of the predictor variables $\{X_k : k = 1, \ldots, K\}$ and the outcome Y:

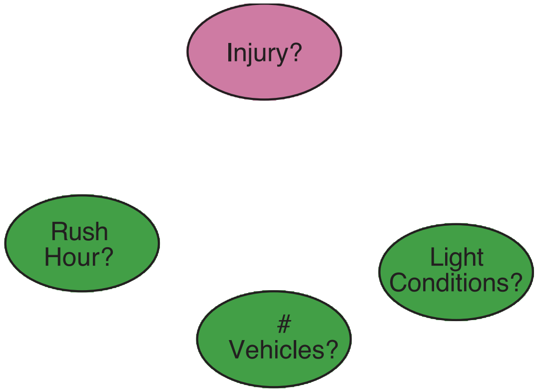
$$Pr(Y, X_1, \ldots, X_K)$$

- Then the conditional probability of Y given the predictors can be calculated as

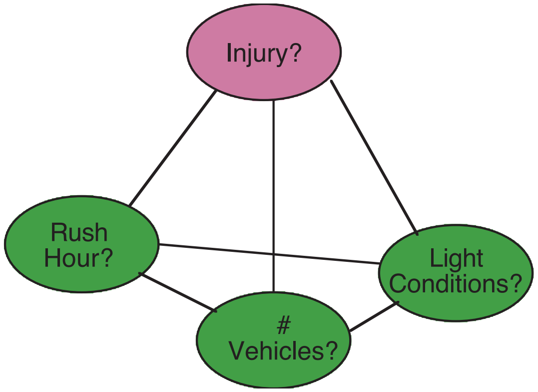
$$\Pr(Y = y|X_1 = x_1, \ldots, X_K = x_K) = \frac{\Pr(Y = y, X_1 = x_1, \ldots, X_K = x_K)}{\sum_{y'} \Pr(Y = y', X_1 = x_1, \ldots, X_K = x_K)}$$

Accidents Variables

How to model $Pr(Y, X_1, \ldots, X_K)$?

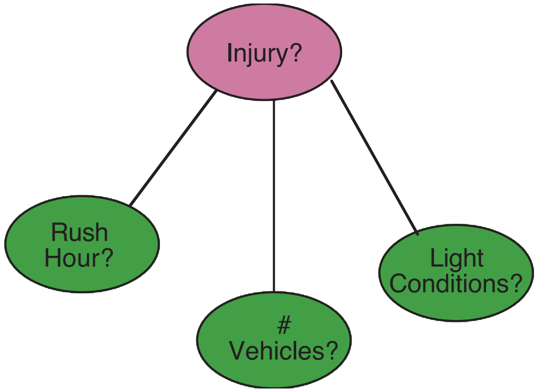


A Full Model



- A full model for $\Pr(Y, X_1, \ldots, X_K)$ must take into account all the two-way, three-way, etc. interactions between the variables!

Naive Bayes



- Naive Bayes only models the interaction between the outcome and each of the predictors.

Naive Bayes

Naive Bayes Assumption:

$$\Pr(X_1, X_2, \ldots, X_K | Y) = \prod_{k=1}^{K} \Pr(X_k | Y)$$

- Naive Bayes assumes that, conditional on the outcome variable, the predictors are independent.
- Do you think this holds for the accidents data?

  A. Yes
  B. No

Naive Bayes

- To fit the model we estimate $\Pr(X_k | Y)$ for each of the predictors
- Then we estimate $\Pr(Y)$
- The joint distribution of Y and the predictors is:

$$\begin{aligned}\Pr(X_1, X_2, \ldots, X_K, Y) &= \Pr(Y)\Pr(X_1, \ldots, X_K | Y) \\ &= \Pr(Y) \prod_{k=1}^{K} \Pr(X_k | Y)\end{aligned}$$

Naive Bayes Training

- We estimate $\Pr(X_k | Y)$ to be equal to the conditional frequency (probability) table from the data
- We estimate $\Pr(Y)$ to be equal to the frequencies from the data

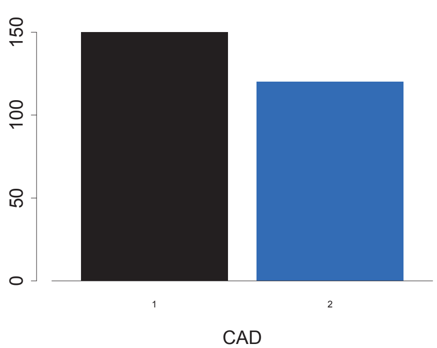
Naive Bayes Training

- Example: $X_1 = \text{Sex}$, $Y = \text{CAD}$
- Need to estimate $Pr(X_1 | Y)$
- Estimate it to be equal to the table from the data:

	Fem	Mal
N	0.45	0.55
Y	0.17	0.83

Naive Bayes Training

- Need to estimate $Pr(Y)$
- Estimate it to be equal to the frequencies from the data:



Naive Bayes Prediction

- Want to be able to tell a patient what is the probability that they have heart disease?
- Want to predict the most probable outcome, i.e. CAD / no CAD
- Want $Pr(Y | X_1, \ldots, X_K)$

Naive Bayes Prediction

- Want $Pr(Y|X_1, \ldots, X_K)$
- We have estimates of $Pr(X_k|Y)$ for each predictor k
- We also have an estimate of $Pr(Y)$

Naive Bayes Prediction

- Naive Bayes Assumption:

$$\mathrm{Pr}(X_1, \ldots, X_K|Y) = \prod_{k=1}^{K} \mathrm{Pr}(X_k|Y)$$

Naive Bayes Prediction

For any values $y, x_1, \ldots, x_K$,

Naive Bayes Prediction

So to predict the probability that $Y = y$, given that $X_1 = x_1, \ldots, X_K = x_K$ we use the formula:

$$Pr(Y = y|X_1 = x_1, \ldots, X_K = x_K)$$
$$= \frac{Pr(Y = y)\left[\prod_{k=1}^{K} Pr(X_k = x_k|Y = y)\right]}{\sum_{y'} Pr(Y = y')\left[\prod_{k=1}^{K} Pr(X_k = x_k|Y = y')\right]}$$

Naive Bayes Prediction

To predict the VALUE of Y, we may take the value y such that $Pr(Y = y|X_1 = x_1, \ldots, X_K = x_K)$ is the largest (e.g., is it more likely that the patient has CAD or does not have CAD?).

Contrast Naive Bayes (which is a model-based classification method) with the heuristic classification tree that we constructed last time. In both cases we used the training data to construct a classification rule, which we then apply to predict Y for new data. What are the differences?

Naive Bayes Prediction

For the CAD example, say we calculate

$$Pr(CAD = \text{Yes}) \left[\prod_{k=1}^{K} Pr(X_k = x_k | CAD = \text{Yes}) \right] = 0.005$$

and

$$Pr(CAD = \text{No}) \left[\prod_{k=1}^{K} Pr(X_k = x_k | CAD = \text{No}) \right] = 0.01$$

Naive Bayes Prediction

What is the conditional probability that CAD = Yes, given the predictor values?

A. In the interval $[0, .25)$

B. In the interval $[.25, .5)$

C. In the interval $[.5, .75)$

D. In the interval $[.75, 1]$

Naive Bayes Prediction

Notice that Y does not have to be a binary variable, as in the CAD example. It can take any finite # of possible values, and the naive Bayes classifier can still be used.