More Naive Bayes

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

Outline

Estimating Conditional Probabilities

- Say the joint counts table for the random variables Z and W in a data set is:

	$W = a$	$W = b$	$W = c$
$Z = 1$	25	13	31
$Z = 2$	73	41	102

Using these data, estimate the conditional distribution of W given that $Z = 2$. Estimate the (marginal) probability distribution of W. Estimate the joint probability that $Z = 1$ and $W = c$. Are Z and W approximately independent?

A. Yes

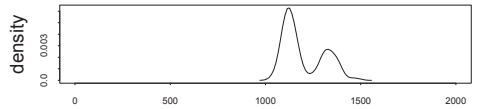
B. No

Classification

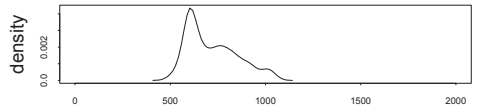
Exploration of some training data gives you the following information: The probability that $Y = 0$ is approximately 52%, and otherwise $Y = 1$. The distribution of X_1 conditional on Y is approximately:

	$X_1 = 1$	$X_1 = 2$
$Y = 0$	0.27	0.73
$Y = 1$	0.65	0.35

The distribution of X_2 conditional on Y is approximately:



X2 Conditioned on Y=0



X2 Conditioned on Y=1

Classification

Say that we want to create a "classification rule" that predicts Y accurately based only on a single predictor variable (either X_1 or X_2). Which predictor variable would you use, to get the best accuracy?

A. X_1

B. X_2

What would be your classification rule? E.g., X_2 is the predictor variable, and the rule is to predict $Y = 1$ when $X_2 \leq 500$ and $Y = 0$ when $X_2 > 500$.

Naive Bayes Training / Prediction

There are 2 steps to a supervised learning method:

1. **TRAINING:** Learn a decision rule (rule for prediction) using a training data set, in which both the predictors and outcome are observed
   -
2. **PREDICTION:** For new observations, predict the value of the outcome variable given the predictors.

Example

Naive Bayes Training:

- For each predictor k estimate the conditional probability table $\Pr(X_k|Y)$ to be equal to the conditional frequency table from the training data, called $\hat{\Pr}(X_k|Y)$
- Estimate $\Pr(Y)$ to be equal to the frequencies from the training data, called $\hat{\Pr}(Y)$

Example

Naive Bayes Prediction:

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

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

Example

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

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

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 $[.75, 1]$

Example

- Example: Y is CAD, predictors X_1 is Sex and X_2 is Exercise Induced Angina
- Estimate $Pr(X_1|Y)$ to be equal to the table from the data:

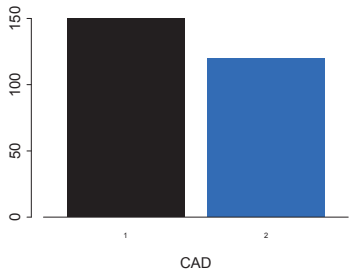
	Sex = Fem	Sex = Mal
CAD = N	0.45	0.55
CAD = Y	0.17	0.83

- Estimate $Pr(X_2|Y)$ to be equal to the table from the data:

	EIA = N	EIA = Y
CAD = N	0.85	0.15
CAD = Y	0.45	0.55

Example

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



- $Pr(Y = \text{Yes}) = 0.56$ and $Pr(Y = \text{No}) = 0.44$

Example

- What is the chance that a female with EIA has CAD?

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

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

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

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

Example

- We need to predict CAD or no CAD for such a patient. What do we pick?
- Typically we predict CAD if $Pr(Y = \text{Yes}|X_1, \ldots, X_K) > 0.5$ and no CAD otherwise
- I.e. we use the threshold 0.5. This is the default choice.
- But we know that not sending a CAD patient to angiography is a more costly mistake than sending a non-CAD patient to angiography
- So maybe we want to use a different threshold (lower or higher?)
- We will discuss this next time

Summary

Naive Bayes is a model-based method. What is the statistical model (assumptions)? What are the parameters of the model? How many parameters are there for the simple CAD example on the previous slides?

- 1-4
- 5-8
- 9-12
- 13-16

Summary

Naive Bayes

For a different problem our training data look like: What are the estimates of $Pr(Y)$, $Pr(X_1|Y)$, and $Pr(X_2|Y)$ for naive Bayes using these data?

X_1	X_2	Y
a	0	1
a	0	0
b	0	1
b	0	1
a	1	1
b	0	0
a	1	1
a	0	0
b	1	0
b	1	0
b	0	1

Naive Bayes

Estimate the probability that $Y = 0$, given that $X_1 = b$ and $X_2 = 0$.

Predict Y, given that $X_1 = b$ and $X_2 = 0$.

Naive Bayes

- Give an example where the naive Bayes assumption is reasonable, i.e., a prediction problem where this assumption would approximately hold.
- Give an example where the naive Bayes assumption definitely would not hold.

Naive Bayes

This assumption does not hold when you are trying to predict the type of aircraft based on its location at a series of time points.

Naive Bayes

In naive Bayes the outcome Y does not have to be a binary variable, as in the CAD example. It can take any finite # of possible values. Examples:

Missing Data

Missing Data in Naive Bayes

- How do we do prediction using naive Bayes if we are missing a value for one of the predictors?

Missing Data

- What do we do if one of the records (rows) in the **test** data is missing a value for one of the predictors?
- When doing classification for that record we simply do not condition on that predictor.
- If we are missing the first predictor X_1, instead of calculating

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

to do classification, we calculate

$$Pr(Y|X_2, \ldots, X_K)$$

- How do we calculate this?

Missing Data

Calculation:

For any possible $y, x_2, \ldots, x_K$,

$$\Pr(Y = y|X_2 = x_2, \ldots, X_K = x_K)$$

$$=$$

$$= \frac{Pr(Y = y) \prod_{k=2}^{K} Pr(X_k = x_k|Y = y)}{\sum_{y'} Pr(Y = y') \prod_{k=2}^{K} Pr(X_k = x_k|Y = y')}$$

Missing Data

- How do we do prediction if there is more than one predictor missing?
- Same: Just leave those predictors out of the product term when calculating the conditional prob. that $Y = y$.

Missing Data

- What do we do if one of the records in the **training** data has a missing value for one of the predictors (say, X_1)?

Missing Data

- When estimating $Pr(X_1|Y)$, we just calculate as if that data point was not in the data set
- How do we change our estimation of $Pr(X_k|Y)$ for $k \neq 1$?

Missing Data

- How do we change our estimation of $Pr(X_k|Y)$ for $k \neq 1$?
- It does not change. Use the whole data set as usual.

Missing Data

Say our training dataset looks like:

X_1	X_2	Y
a	1	1
a	0	0
b	0	1
b	0	1
a	1	1
b	1	0
?	1	1
?	0	0
b	1	0
b	1	0
b	0	1

Missing Data

What are the (frequentist) estimates of $Pr(Y)$, $Pr(X_1|Y)$, and $Pr(X_2|Y)$ for naive Bayes?

X_1	X_2	Y
a	1	1
a	0	0
b	0	1
b	0	1
a	1	1
b	1	0
?	1	1
?	0	0
b	1	0
b	1	0
b	0	1

Missing Data

- Estimate the probability that $Y = 1$, given that $X_1 = b$ and $X_2 = 0$.

- Using threshold 0.5, predict the value of Y, knowing that $X_1 = b$ and $X_2 = 0$.

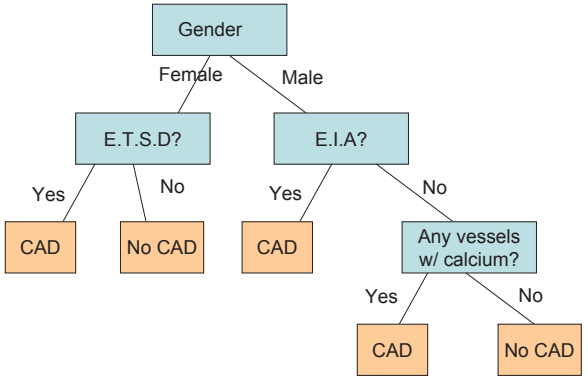
Missing Data

- Estimate the probability that $Y = 1$, given that $X_2 = 0$ and X_1 is missing.

- Using threshold 0.4, predict the value of Y, knowing that $X_2 = 0$ and X_1 is missing.

Naive Bayes on the Heart Disease Data

- Recall that we constructed a classification tree for heart disease using exploratory data analysis:



Naive Bayes on the Heart Disease Data

- This tree had 80% accuracy on the training data
- I never told you the accuracy on the test data

Naive Bayes on the Heart Disease Data

The training data (from Cleveland):

```
63.0,1.0,1.0,145.0,233.0,1.0,2.0,150.0,0.0,2.3,3.0,0.0,6.0,0
67.0,1.0,4.0,160.0,286.0,0.0,2.0,108.0,1.0,1.5,2.0,3.0,3.0,2
67.0,1.0,4.0,120.0,229.0,0.0,2.0,129.0,1.0,2.6,2.0,2.0,7.0,1
37.0,1.0,3.0,130.0,250.0,0.0,0.0,187.0,0.0,3.5,3.0,0.0,3.0,0
41.0,0.0,2.0,130.0,204.0,0.0,2.0,172.0,0.0,1.4,1.0,0.0,3.0,0
56.0,1.0,2.0,120.0,236.0,0.0,0.0,178.0,0.0,0.8,1.0,0.0,3.0,0
62.0,0.0,4.0,140.0,268.0,0.0,2.0,160.0,0.0,3.6,3.0,2.0,3.0,3
57.0,0.0,4.0,120.0,354.0,0.0,0.0,163.0,1.0,0.6,1.0,0.0,3.0,0
63.0,1.0,4.0,130.0,254.0,0.0,2.0,147.0,0.0,1.4,2.0,1.0,7.0,2
53.0,1.0,4.0,140.0,203.0,1.0,2.0,155.0,1.0,3.1,3.0,0.0,7.0,1
57.0,1.0,4.0,140.0,192.0,0.0,0.0,148.0,0.0,0.4,2.0,0.0,6.0,0
56.0,0.0,2.0,140.0,294.0,0.0,2.0,153.0,0.0,1.3,2.0,0.0,3.0,0
56.0,1.0,3.0,130.0,256.0,1.0,2.0,142.0,1.0,0.6,2.0,1.0,6.0,2
44.0,1.0,2.0,120.0,263.0,0.0,0.0,173.0,0.0,0.0,1.0,0.0,7.0,0
52.0,1.0,3.0,172.0,199.0,1.0,0.0,162.0,0.0,0.5,1.0,0.0,7.0,0
```


Naive Bayes on the Heart Disease Data

The last column is heart disease status (0-3)

```
63.0,1.0,1.0,145.0,233.0,1.0,2.0,150.0,0.0,2.3,3.0,0.0,6.0,0
67.0,1.0,4.0,160.0,286.0,0.0,2.0,108.0,1.0,1.5,2.0,3.0,3.0,2
67.0,1.0,4.0,120.0,229.0,0.0,2.0,129.0,1.0,2.6,2.0,2.0,7.0,1
37.0,1.0,3.0,130.0,250.0,0.0,0.0,187.0,0.0,3.5,3.0,0.0,3.0,0
41.0,0.0,2.0,130.0,204.0,0.0,2.0,172.0,0.0,1.4,1.0,0.0,3.0,0
56.0,1.0,2.0,120.0,236.0,0.0,0.0,178.0,0.0,0.8,1.0,0.0,3.0,0
62.0,0.0,4.0,140.0,268.0,0.0,2.0,160.0,0.0,3.6,3.0,2.0,3.0,3
57.0,0.0,4.0,120.0,354.0,0.0,0.0,163.0,1.0,0.6,1.0,0.0,3.0,0
63.0,1.0,4.0,130.0,254.0,0.0,2.0,147.0,0.0,1.4,2.0,1.0,7.0,2
53.0,1.0,4.0,140.0,203.0,1.0,2.0,155.0,1.0,3.1,3.0,0.0,7.0,1
57.0,1.0,4.0,140.0,192.0,0.0,0.0,148.0,0.0,0.4,2.0,0.0,6.0,0
56.0,0.0,2.0,140.0,294.0,0.0,2.0,153.0,0.0,1.3,2.0,0.0,3.0,0
56.0,1.0,3.0,130.0,256.0,1.0,2.0,142.0,1.0,0.6,2.0,1.0,6.0,2
44.0,1.0,2.0,120.0,263.0,0.0,0.0,173.0,0.0,0.0,1.0,0.0,7.0,0
52.0,1.0,3.0,172.0,199.0,1.0,0.0,162.0,0.0,0.5,1.0,0.0,7.0,0
```


Naive Bayes on the Heart Disease Data

We just predicted heart disease present (1-3) or absent (0)

```
63.0,1.0,1.0,145.0,233.0,1.0,2.0,150.0,0.0,2.3,3.0,0.0,6.0,0
67.0,1.0,4.0,160.0,286.0,0.0,2.0,108.0,1.0,1.5,2.0,3.0,3.0,2
67.0,1.0,4.0,120.0,229.0,0.0,2.0,129.0,1.0,2.6,2.0,2.0,7.0,1
37.0,1.0,3.0,130.0,250.0,0.0,0.0,187.0,0.0,3.5,3.0,0.0,3.0,0
41.0,0.0,2.0,130.0,204.0,0.0,2.0,172.0,0.0,1.4,1.0,0.0,3.0,0
56.0,1.0,2.0,120.0,236.0,0.0,0.0,178.0,0.0,0.8,1.0,0.0,3.0,0
62.0,0.0,4.0,140.0,268.0,0.0,2.0,160.0,0.0,3.6,3.0,2.0,3.0,3
57.0,0.0,4.0,120.0,354.0,0.0,0.0,163.0,1.0,0.6,1.0,0.0,3.0,0
63.0,1.0,4.0,130.0,254.0,0.0,2.0,147.0,0.0,1.4,2.0,1.0,7.0,2
53.0,1.0,4.0,140.0,203.0,1.0,2.0,155.0,1.0,3.1,3.0,0.0,7.0,1
57.0,1.0,4.0,140.0,192.0,0.0,0.0,148.0,0.0,0.4,2.0,0.0,6.0,0
56.0,0.0,2.0,140.0,294.0,0.0,2.0,153.0,0.0,1.3,2.0,0.0,3.0,0
56.0,1.0,3.0,130.0,256.0,1.0,2.0,142.0,1.0,0.6,2.0,1.0,6.0,2
44.0,1.0,2.0,120.0,263.0,0.0,0.0,173.0,0.0,0.0,1.0,0.0,7.0,0
52.0,1.0,3.0,172.0,199.0,1.0,0.0,162.0,0.0,0.5,1.0,0.0,7.0,0
```

Naive Bayes on the Heart Disease Data

The other columns are age, gender (0/1), E.I.A. (0/1), etc.

```
63.0,1.0,1.0,145.0,233.0,1.0,2.0,150.0,0.0,2.3,3.0,0.0,6.0,0
67.0,1.0,4.0,160.0,286.0,0.0,2.0,108.0,1.0,1.5,2.0,3.0,3.0,2
67.0,1.0,4.0,120.0,229.0,0.0,2.0,129.0,1.0,2.6,2.0,2.0,7.0,1
37.0,1.0,3.0,130.0,250.0,0.0,0.0,187.0,0.0,3.5,3.0,0.0,3.0,0
41.0,0.0,2.0,130.0,204.0,0.0,2.0,172.0,0.0,1.4,1.0,0.0,3.0,0
56.0,1.0,2.0,120.0,236.0,0.0,0.0,178.0,0.0,0.8,1.0,0.0,3.0,0
62.0,0.0,4.0,140.0,268.0,0.0,2.0,160.0,0.0,3.6,3.0,2.0,3.0,3
57.0,0.0,4.0,120.0,354.0,0.0,0.0,163.0,1.0,0.6,1.0,0.0,3.0,0
63.0,1.0,4.0,130.0,254.0,0.0,2.0,147.0,0.0,1.4,2.0,1.0,7.0,2
53.0,1.0,4.0,140.0,203.0,1.0,2.0,155.0,1.0,3.1,3.0,0.0,7.0,1
57.0,1.0,4.0,140.0,192.0,0.0,0.0,148.0,0.0,0.4,2.0,0.0,6.0,0
56.0,0.0,2.0,140.0,294.0,0.0,2.0,153.0,0.0,1.3,2.0,0.0,3.0,0
56.0,1.0,3.0,130.0,256.0,1.0,2.0,142.0,1.0,0.6,2.0,1.0,6.0,2
44.0,1.0,2.0,120.0,263.0,0.0,0.0,173.0,0.0,0.0,1.0,0.0,7.0,0
```


Naive Bayes on the Heart Disease Data

The test data (from California):

```
63,1,4,140,260,0,1,112,1,3,2,?,?,2
44,1,4,130,209,0,1,127,0,0,?,?,?,0
60,1,4,132,218,0,1,140,1,1.5,3,?,?,2
55,1,4,142,228,0,1,149,1,2.5,1,?,?,1
66,1,3,110,213,1,2,99,1,1.3,2,?,?,0
66,1,3,120,0,0,1,120,0,-0.5,1,?,?,0
65,1,4,150,236,1,1,105,1,0,?,?,?,3
60,1,3,180,0,0,1,140,1,1.5,2,?,?,0
60,1,3,120,0,?,0,141,1,2,1,?,?,3
60,1,2,160,267,1,1,157,0,0.5,2,?,?,1
56,1,2,126,166,0,1,140,0,0,?,?,?,0
59,1,4,140,0,0,1,117,1,1,2,?,?,1
62,1,4,110,0,0,0,120,1,0.5,2,?,3,1
63,1,3,?,0,0,2,?,?,?,?,?,?,1
```


Naive Bayes on the Heart Disease Data

The last column is the heart disease status (0-3)

```
63,1,4,140,260,0,1,112,1,3,2,?,?,2
44,1,4,130,209,0,1,127,0,0,?,?,?,0
60,1,4,132,218,0,1,140,1,1.5,3,?,?,2
55,1,4,142,228,0,1,149,1,2.5,1,?,?,1
66,1,3,110,213,1,2,99,1,1.3,2,?,?,0
66,1,3,120,0,0,1,120,0,-0.5,1,?,?,0
65,1,4,150,236,1,1,105,1,0,?,?,?,3
60,1,3,180,0,0,1,140,1,1.5,2,?,?,0
60,1,3,120,0,?,0,141,1,2,1,?,?,3
60,1,2,160,267,1,1,157,0,0.5,2,?,?,1
56,1,2,126,166,0,1,140,0,0,?,?,?,0
59,1,4,140,0,0,1,117,1,1,2,?,?,1
62,1,4,110,0,0,0,120,1,0.5,2,?,3,1
63,1,3,?,0,0,2,?,?,?,?,?,?,1
```


Naive Bayes on the Heart Disease Data

The test data has a lot of missing values for the predictors!

```
63,1,4,140,260,0,1,112,1,3,2,?,?,2
44,1,4,130,209,0,1,127,0,0,?,?,?,0
60,1,4,132,218,0,1,140,1,1.5,3,?,?,2
55,1,4,142,228,0,1,149,1,2.5,1,?,?,1
66,1,3,110,213,1,2,99,1,1.3,2,?,?,0
66,1,3,120,0,0,1,120,0,-0.5,1,?,?,0
65,1,4,150,236,1,1,105,1,0,?,?,?,3
60,1,3,180,0,0,1,140,1,1.5,2,?,?,0
60,1,3,120,0,?,0,141,1,2,1,?,?,3
60,1,2,160,267,1,1,157,0,0.5,2,?,?,1
56,1,2,126,166,0,1,140,0,0,?,?,?,0
59,1,4,140,0,0,1,117,1,1,2,?,?,1
62,1,4,110,0,0,0,120,1,0.5,2,?,3,1
63,1,3,?,0,0,2,?,?,?,?,?,?,1
```

Naive Bayes on the Heart Disease Data

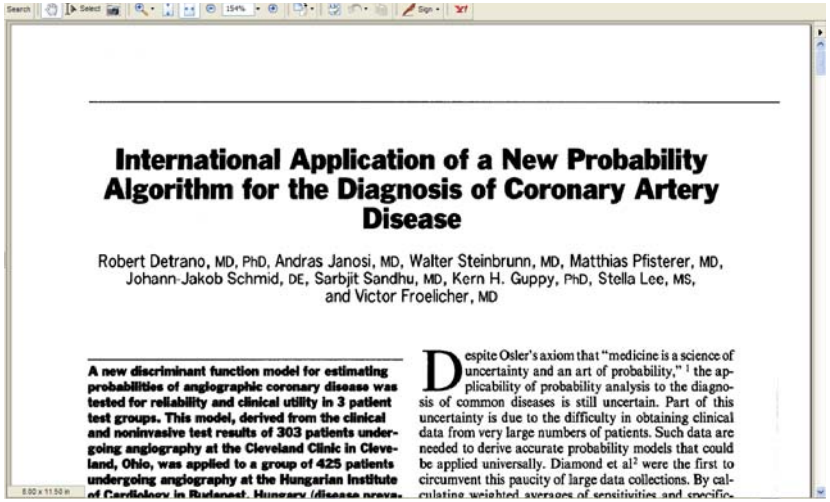
- We can't use our classification tree when we are missing the value of one of the predictors that is used in the tree.
- In fact, the original heart disease training data (303 patients) had missing data as well
- The missing data had been removed in the data set that we used in Lecture 3, leaving 270 patients.
- But now we can fit a naive Bayes classifier to the full training data, and predict on the test data, regardless of the missing values!

Naive Bayes on the Heart Disease Data

- Applying naive Bayes to the heart disease data gives a 75% accuracy on the test data.

Naive Bayes on the Heart Disease Data

- Recall the article on heart disease prediction:



International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease

Robert Detrano, MD, PhD, Andras Janosi, MD, Walter Steinbrunn, MD, Matthias Pfisterer, MD, Johann-Jakob Schmid, DE, Sarbjit Sandhu, MD, Kern H. Guppy, PhD, Stella Lee, MS, and Victor Froelicher, MD

A new discriminant function model for estimating probabilities of angiographic coronary disease was tested for reliability and clinical utility in 3 patient test groups. This model, derived from the clinical and noninvasive test results of 303 patients undergoing angiography at the Cleveland Clinic in Cleveland, Ohio, was applied to a group of 425 patients undergoing angiography at the Hungarian Institute

Despite Osler's axiom that "medicine is a science of uncertainty and an art of probability," [1] the applicability of probability analysis to the diagnosis of common diseases is still uncertain. Part of this uncertainty is due to the difficulty in obtaining clinical data from very large numbers of patients. Such data are needed to derive accurate probability models that could be applied universally. Diamond et al [2] were the first to circumvent this paucity of large data collections. By calculating weighted averages of sensitivities and specific-

Naive Bayes on the Heart Disease Data

- The authors report an accuracy of about 77% for their methods on the same test data.
- Our classifier, despite being "naive", does almost as well!