- Probability Primer
- Naïve Bayes
  - Bayes' Rule
  - Conditional Probabilities
  - Probabilistic Models

Motivation

- In many datasets, relationship between attributes and a class variable is <u>non-deterministic</u>.
- Why?
  - Noisy data
  - Confounding and interaction of factors
  - Relevant variables not included in the data
- *Scenario*:
  - Risk of heart disease based on individual's diet and workout frequency

Scenario

- Risk of heart disease based on individual’s diet and workout frequency
  - Most people who “work out” and have a healthy diet don’t get heart disease
  - Yet, some healthy individuals still do:
    - Smoking, alcohol abuse, …

What we're trying to do

- Model probabilistic relationships
- *"What is the probability that this person will get heart disease, given their diet and workout regimen?"*
  - Output is most similar to Logistic Regression
- Will introduce naïve Bayes model
  - A type of Bayesian classifier
  - *More advanced:* Bayesian network

Bayes Classifier

- A probabilistic framework for solving *classification* problems
  - Used in both naïve Bayes and Bayesian networks
- Based on Bayes' Theorem:

What's the probability that it rains today *AND* that I'm carrying an umbrella?

Terminology/Notation Primer

- X and Y (two different variables)
- Joint probability: $P(X=x, Y=y)$
  - The probability that variable X takes on the value x and variable Y has the value y
- Conditional probability: $P(Y=y \mid X=x)$
  - Probability that variable Y has the value y, given that variable X takes on the value x

Given that I'm observed with an umbrella, what's the probability that it will rain today?

Terminology/Notation Primer

- Single Probability: $P(X = x)$
  - "X has the value x"
- Joint Probability: $P(X, Y)$
  - "X and Y"
- Conditional Probability: $P(Y \mid X)$
  - "Y" given observation of "X"
- Relation of Joint and Conditional Probabilities:

$$P(X, Y) = P(Y \mid X) \times P(X)$$

Terminology/Notation Primer

$$P(X,Y) = P(Y \mid X) \times P(X)$$

$$P(Y,X) = P(X \mid Y) \times P(Y)$$

$$P(X,Y) = P(Y,X)$$

$$P(X,Y) = P(Y \mid X) \times P(X) = P(X \mid Y) \times P(Y)$$

<u>Bayes' Theorem:</u> $$P(Y \mid X) = \frac{P(X \mid Y) \times P(Y)}{P(X)}$$

Predicted Probability Example

- Scenario:
  1. A doctor knows that meningitis causes a stiff neck 50% of the time
  2. <u>Prior probability</u> of any patient having meningitis is 1/50,000
  3. <u>Prior probability</u> of any patient having a stiff neck is 1/20
- If a patient has a stiff neck, what's the probability that they have meningitis?

- A doctor knows that meningitis causes a stiff neck 50% of the time
- Prior probability of any patient having meningitis is 1/50,000
- Prior probability of any patient having a stiff neck is 1/20

- Apply Bayes' Rule:
- If a patient has a stiff neck, what's the probability that they have meningitis?
  - Interested in: $P(M \mid S)$

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)}$$

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times \frac{1}{50000}}{\frac{1}{20}} = 0.0002$$

Very low probability

How to Apply Bayes' Theorem to Data Mining and Datasets?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- *Target class: Evade*
- *Predictor variables: Refund, Status, Income*
- What is probability of *Evade* given the values of *Refund, Status, Income?*

$$P(E \mid R, S, I)$$

Above .5? Predict YES, else predict NO.

How to Apply Bayes' Theorem to Data Mining and Datasets?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- How to compute? $P(E \mid R, S, I)$
- Need test instance:
  - What are values of *R*, *S*, *I*?
  - Test instance is:
    - Refund=Yes
    - Status=Married
    - Income=60K

Issue: we don't have any training example that these same three attributes values.

Naïve Bayes Classifier

- Why called naïve?
  - Assumes that attributes (predictor variables) are <u>conditionally independent</u>.
    - No correlation
  - Big assumption!
- What is conditionally independent?
  - Variable X is conditionally independent of Y if the following holds:

$$P(X \mid Y, Z) = P(X \mid Z)$$

Conditional Independence

Assuming variables *X* and *Y* and conditionally independent, can derive:

"given Z, what is the joint probability of X and Y?"

$$\boxed{P(X,Y \mid Z)} = \frac{P(X,Y,Z)}{P(Z)}$$

$$= \frac{P(X,Y,Z)}{P(Y,Z)} \times \frac{P(Y,Z)}{P(Z)}$$

$$= P(X \mid Y,Z) \times P(Y \mid Z)$$

$$= \boxed{P(X \mid Z) \times P(Y \mid Z)}$$

Naïve Bayes Classifier

Before (simple Bayes' rule):
Single predictor variable X

$$P(Y \mid X) = \frac{P(X \mid Y) \times P(Y)}{P(X)}$$

Now we have a bunch of predictor
Variables: $X_1, X_2, X_3, \ldots, X_n$

$$P(Y \mid X_1, X_2, \ldots, X_n)$$

$$P(Y \mid X_1, X_2, \ldots, X_n) = \frac{P(X_1, X_2, \ldots, X_n \mid Y) \times P(Y)}{P(X_1, X_2, \ldots, X_n)}$$

$$P(Y \mid X) = \frac{P(Y) \prod_{i=1}^{d} P(X_i \mid Y)}{P(X_1, X_2, \ldots, X_n)}$$

Naïve Bayes Classifier

$$P(Y \mid X) = \frac{P(Y)\prod_{i=1}^{d} P(X_i \mid Y)}{P(X_1, X_2, \ldots, X_n)}$$

For binary problems: $P(Y|X) > .5$? Predict YES, else predict NO.

Example: will compute *P(E=Yes | Status, Income, Refund)* and *P(E=No | Status, Income, Refund)*

- Find which one is greater (greater likelihood)

Can compute from training data:

$P(E = Yes)$ $P(E = No)$ $P(Y)$

$P(\text{Refund} = No \mid E = Yes)$

$P(X_1 \mid Y)$ $P(X_3 \mid Y)$

Cannot compute / hard to compute:

$$P(X_1, X_2, \ldots, X_n)$$

$$P(\text{Refund}=yes, \text{Status}=married, \text{Income}=120k)$$

- Not a problem, since the two denominators will be the same.
- Need to see which numerator is greater.

Estimating Prior Probabilities for the Class target $P(Y)$

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

- *P(Evade=yes)*
- *= 3/10*
- *P(Evade=no)*
- *=7/10*

Estimating Conditional Probabilities for Categorical Attributes $P(X_1 | Y)$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- $P(Refund=yes | Evade=no)$
- $= 3/7$
- $P(Status=married | Evade=yes)$
- $=0/3$
  - *Yikes!*
  - *Will handle the 0% probability later*

Estimating Conditional Probabilities for Continuous Attributes $P(X_1 | Y)$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- For continuous attributes:
  1. Discretize into bins
  2. Two-way split:
     - $(A <= v)$ or $(A > v)$

Full Example

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 75\text{K})$$

P(NO) = 7/10
P(YES) = 3/10

P(Refund=YES|NO) = 3/7
P(Refund=NO|NO) = 4/7
P(Refund=YES|YES) = 0/3
P(Refund=NO|YES) = 3/3

P(Status=SINGLE|NO) = 2/7
P(Status=DIVORSED|NO) = 1/7
P(Status=MARRIED|NO) = 4/7
P(Status=SINGLE|YES) = 2/3
P(Status=DIVORSED|YES) = 1/3
P(Status=MARRIED|YES) = 0/3

For taxable income:
P(Income=above 101K|NO) = 3/7
P(Income=below101K|NO) = 4/7
P(Income=above 101K|YES) = 0/3
P(Income=below 101K|YES) = 3/3

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 75\text{K})$$

P(NO) = 7/10
P(YES) = 3/10

P(Refund=YES|NO) = 3/7
P(Refund=NO|NO) = 4/7
P(Refund=YES|YES) = 0/3
P(Refund=NO|YES) = 3/3

P(Status=SINGLE|NO) = 2/7
P(Status=DIVORSED|NO) = 1/7
P(Status=MARRIED|NO) = 4/7
P(Status=SINGLE|YES) = 2/3
P(Status=DIVORSED|YES) = 1/3
P(Status=MARRIED|YES) = 0/3

For taxable income:
P(Income=above 101K|NO) = 3/7
P(Income=below101K|NO) = 4/7
P(Income=above 101K|YES) = 0/3
P(Income=below 101K|YES) = 3/3

P(X|Class=No) = P(Refund=No|Class=No)
× P(Married| Class=No)
× P(Income=below 101K| Class=No)
= 4/7 × 4/7 × 4/7 = 0.1866

P(X|Class=Yes) = P(Refund=No| Class=Yes)
× P(Married| Class=Yes)
× P(Income=below 101K| Class=Yes)
= 1 × 0 × 1 = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)

=> Class = No

Smoothing of Conditional Probabilities

- If one of the conditional probabilities is 0, then the entire product will be 0
- *Idea:* Instead use very small non-zeros values, such as 0.00001

$$\text{Original: } P(x_i \mid y_j) = \frac{n_c}{n}$$

$$n = \text{\# of training examples that have value } y_j$$

$$n_c = \text{\# of examples from class } y_j \text{ that take on value } x_i$$

Smoothing of Conditional Probabilities

- *Idea:* Instead use very small non-zeros values, such as 0.00001

$$\text{Original: } P(x_i \mid y_j) = \frac{n_c}{n}$$

$$\text{Laplace: } P(x_i \mid y_j) = \frac{n_c + 1}{n + C}$$

$n = \#$ of training examples that have value y_j

$n_c = \#$ of examples from class y_j that take on value x_i

$C = \#$ of classes

w/ Laplace Smoothing:

P(NO) = 7/10
P(YES) = 3/10

P(Refund=YES|NO) = 4/9
P(Refund=NO|NO) = 5/9
P(Refund=YES|YES) = 1/5
P(Refund=NO|YES) = 4/5

P(Status=SINGLE|NO) = 3/9
P(Status=DIVORSED|NO) = 2/9
P(Status=MARRIED|NO) = 5/9
P(Status=SINGLE|YES) = 3/5
P(Status=DIVORSED|YES) = 2/5
P(Status=MARRIED|YES) = 1/5

For taxable income:
P(Income=above 101K|NO) = 4/9
P(Income=below101K|NO) = 5/9
P(Income=above 101K|YES) = 1/5
P(Income=below 101K|YES) = 4/5

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 75\text{K})$$

- P(X|Class=No) = P(Refund=No|Class=No)
  × P(Married| Class=No)
  × P(Income=below 101K| Class=No)
  = 5/9 × 5/9 × 5/9 = 0.1715

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=below 101K| Class=Yes)
  = 4/5 × 1/5 × 4/5 = 0.128

Is P(X|No)P(No) > P(X|Yes)P(Yes)?

.1715 x 7/10 > .128 x 3/10

Therefore P(No|X) > P(Yes|X)
=> Class = No

Characteristics of Naïve Bayes Classifiers

- Robust to isolated noise
  - Noise is averaged out by estimating the conditional probabilities from data
- Handling missing values
  - Simply ignore them when estimating the probabilities
- Robust to irrelevant attributes
  - If X_i is an irrelevant attribute, then $P(X_i \mid Y)$ becomes almost uniformly distributed
    - P(Refund=Yes|YES)=0.5
    - P(Refund=Yes|NO)=0.5

Characteristics of Naïve Bayes Classifiers

- Independence assumption may not hold for some attributes
  - Correlated attributes can degrade performance of naïve Bayes
- But … naïve Bayes (for such a simple model), still works surprisingly well even when there is some correlation between attributes

References

- *Fundamentals of Machine Learning for Predictive Data Analytics*, 1st Edition, Kelleher et al.
- *Introduction to Data Mining*, 1st edition, Tam et al.
- *Data Mining and Business Analytics with R*, 1st edition, Ledolter