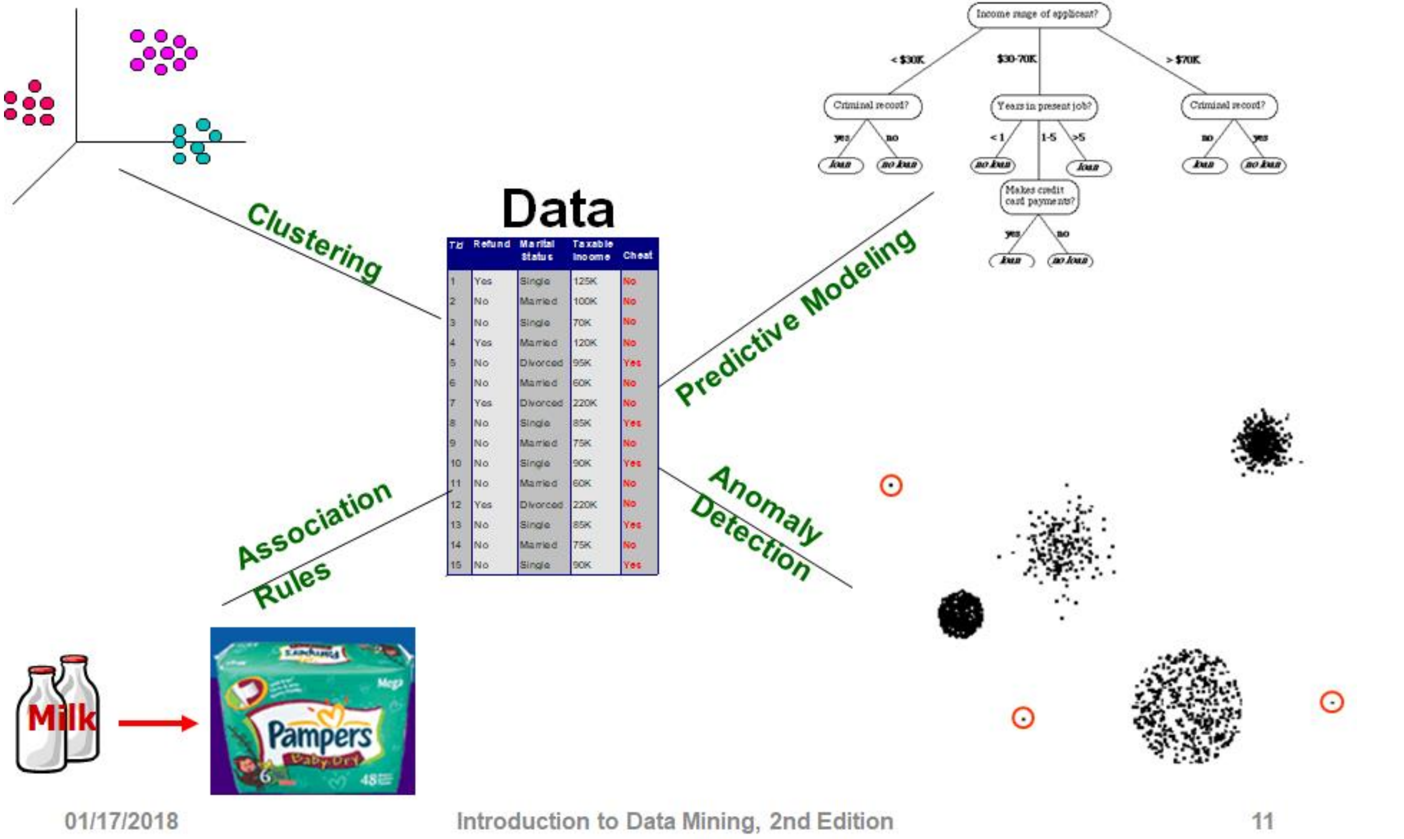
Foundations of Data Science & Analytics: Naive Bayes Classifier

Ezgi Siir Kibris

Introduction to Data Mining, 2nd Edition
by
Tan, Steinbach, Karpatne, Kumar

Tasks



Tid	Refund	Marital Status	Taxable Income	Cheat
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
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Predictive Modeling

	Output
Classification:	Classes / Categories
Regression:	Continuous Values

Classification: Definition

- Given a collection of data instances (training set)
  - Each instance is characterized by a tuple (*x*, *y*), where *x* is the feature set and *y* is the class label
    - *x*: feature, attribute, independent variable, input
    - *y*: class, response, dependent variable, output
- Task:
  - Learn a model that maps each feature set *x* into one of the predefined class labels *y*

x: feature set

y: class

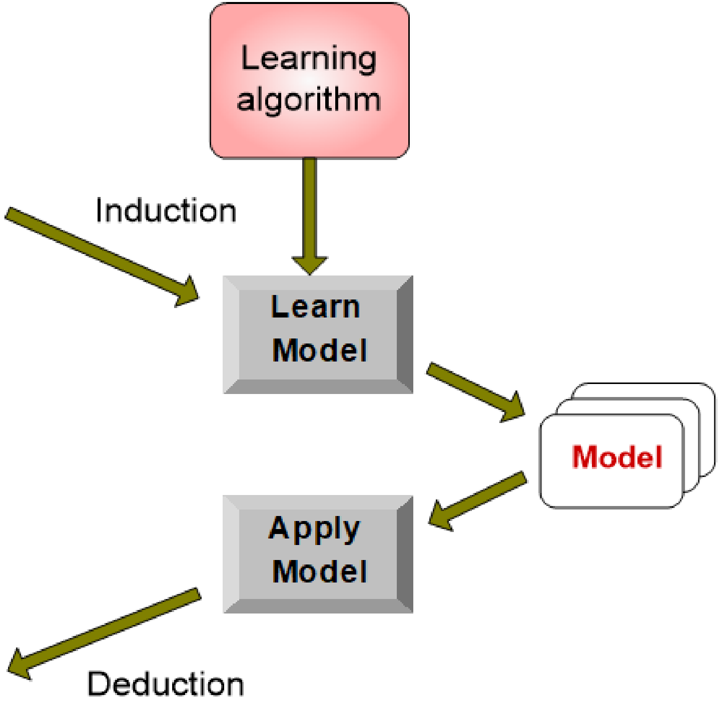
Cat
Dog
Bird
Lion
...

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Medium	100K	?
12	Yes	Medium	80K	?
13	No	Small	95K	?

Test Set

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Medium	100K	?
12	Yes	Medium	80K	?
13	No	Small	95K	?

Test Set

We found the exact same data instance in the training set, so we classify Instance 11 as “No”. **Classification is easy?!**

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Medium	100K	?
12	Yes	Medium	80K	?
13	No	Small	95K	?

Test Set

We don’t see an exact matching data instance in the training set. The closest is Instance 4, but that instance’s Attrib3 value is much larger than Instance 12.

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Medium	100K	?
12	Yes	Medium	80K	?
13	No	Small	95K	?

Test Set

No exactly matching instance but Instance 10 is very close, as is Instance 8. Both of those have a class value of “Yes”, so maybe it’s ok to classify Instance 13 as “Yes”?

Classification Techniques

- **Base Classifiers**
  - Decision Tree based Methods
  - Rule-based Methods
  - Instance-based Methods (Nearest-Neighbor)
  - **Naïve Bayes**
  - Support Vector Machines
  - Neural Networks and Deep Learning
- **Ensemble Classifiers**
  - Boosting, Bagging, Random Forests

Bayes Classifier

- **Conditional Probability:**

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

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

- **Bayes' theorem:**

Likelihood

Posterior probability

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

Class prior probability

Bayes Classifier

- **Bayes theorem:**

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

- **We want to choose Y that is maximal**
  P(Y | X1, X2, …, Xd)

- **Equivalent to maximizing**
  P(X1, X2, …, Xd | Y) P(Y)

Naive Bayes Independence Assumption

Assume conditional independence among features X_i when class is given:

P(X1, X2, …, Xd | Y) = P(X1 | Y) P(X2 | Y)… P(Xd | 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

- **Given a test data point:**

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

- **Can we estimate:**
  P(Evade=Yes | X) and
  P(Evade=No | X)?

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

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

- **Maximize:**
  P(X1, X2, …, Xd|Y) P(Y)

- **P(Y):**
  P(Evade=Yes) =
  P(Evade=No) =

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

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

- **Maximize:**
  P(X1, X2, …, Xd | Y) P(Y)

- **P(X1, X2, …, Xd | Y):**
  How to compute?

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

Now:

P(Refund=No, Divorced, Income=120K | Yes)

= P(Refund=No | Yes) P(Divorced | Yes) P(Income=120K | Yes)

= 3/3 * 1/3 * P(Income=120K | Yes)

Continuous Features

- We cannot directly use continuous features with Naïve Bayes. The easiest way to deal with continuous features is to discretize them!

Discretize the Continuous Feature

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	High	No
2	No	Married	High	No
3	No	Single	Med	No
4	Yes	Married	High	No
5	No	Divorced	Med	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Med	Yes
9	No	Married	Med	No
10	No	Single	Med	Yes

Basic discretization scheme for Taxable Income

- [100K, 300K] = High
- [70K, 100K) = Med
- (0K, 70K) = Low

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	High	No
2	No	Married	High	No
3	No	Single	Med	No
4	Yes	Married	High	No
5	No	Divorced	Med	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Med	Yes
9	No	Married	Med	No
10	No	Single	Med	Yes

P(Refund=No, Divorced, Income=90K | Yes)

= P(Refund=No, Divorced, Income=Med | Yes)

= P(No | Yes) * (Divorced | Yes) * (Med | Yes)

= 3/3 * 1/3 * 3/3 = 0.333

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	High	No
2	No	Married	High	No
3	No	Single	Med	No
4	Yes	Married	High	No
5	No	Divorced	Med	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Med	Yes
9	No	Married	Med	No
10	No	Single	Med	Yes

P(Refund=No, Divorced, Income=Med | Yes)

= 3/3 * 1/3 * 3/3 = 0.333

P(Refund=No, Divorced, Income=Med | No)

= P(No | No) * P(Divorced | No) * P(Med | No)

= 4/7 * 1/7 * 2/7 = 0.0233

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	High	No
2	No	Married	High	No
3	No	Single	Med	No
4	Yes	Married	High	No
5	No	Divorced	Med	Yes
6	No	Married	Low	No
7	Yes	Divorced	High	No
8	No	Single	Med	Yes
9	No	Married	Med	No
10	No	Single	Med	Yes

P(Yes | X) = P(X | Yes) * P(Yes)
= 0.333 * 3/10 = 0.0999

P(No | X) = P(X | No) * P(No)
= 0.0233 * 7/10 = 0.01631

Predict:
Evade = Yes
Probability = P(Yes | X) / (P(Yes | X) + P(No | X))
= 0.0999 / (0.0999 + 0.01631)
= 0.860 = 86.0%

Major Issue with Naive Bayes

If one of the conditional probabilities is zero, then the entire expression becomes zero

original: $P(X_i = c|y) = \dfrac{n_c}{n}$

Laplace Estimate: $P(X_i = c|y) = \dfrac{n_c + 1}{n + v}$

n: number of training instances belonging to class y

n_c: number of instances with $X_i = c$ and $Y = y$

v: total number of attribute values that X_i can take

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

Original

$$P(A \mid M) =$$

$$P(A \mid N) =$$

$$P(A \mid M)P(M) =$$

$$P(A \mid N)P(N) =$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

$$P(M|A) = \frac{P(A|M)P(M)}{P(A|M)P(M)+P(A|N)P(N)} = 0.886$$

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Laplace Estimate: $P(X_i = c|y) = \frac{n_c + 1}{n + v}$

$$P(A|M) = \frac{7}{9} \times \frac{7}{9} \times \frac{3}{10} \times \frac{3}{9} = 0.06$$

$$P(A|N) = \frac{2}{15} \times \frac{11}{15} \times \frac{4}{16} \times \frac{5}{15} = 0.008$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.008 \times \frac{13}{20} = 0.005$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

$$P(M|A) = \frac{P(A|M)P(M)}{P(A|M)P(M)+P(A|N)P(N)} = 0.808$$

Naïve Bayes easily handles missing values

Predict for

X = (Refund = Yes, Divorced, Income = ?)

$$P(X|No) = 3/6 \times 1/7$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	?	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

Training a Naive Bayes Model

- **Pre-calculate:**
  - P(y) for y in Y (these are our ”prior” probabilities)
  - $P(X_i \mid y)$ for y in Y, for X_i in X
    - If X_i is discrete, pre-calculate $P(x_i \mid y)$ for x_i in X_i
    - Else if X_i is continuous, discretize the feature to create categories, then pre-calculate $P(x_i \mid y)$ for x_i in X_i

Assignment 4