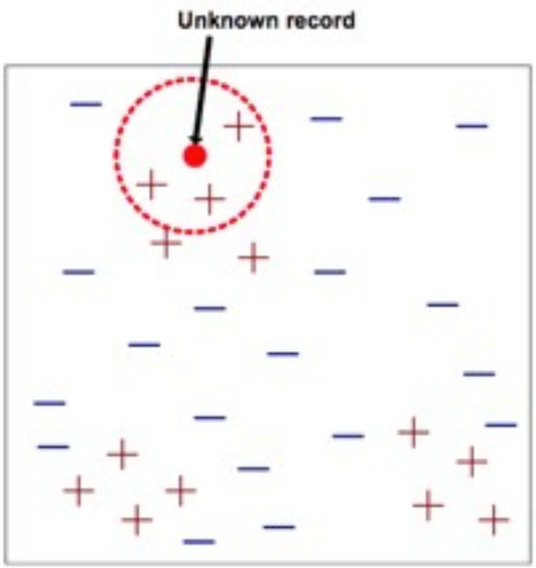
Data Mining
Classification: Alternative Techniques

Mauro Sozio

slides adapted from “Introduction to Data Mining” by Tan, Steinbach, Kumar.

K-Nearest Neighbor Classifier

- Training:
  - turn dataset into vectors (e.g. points euclidean space)
  - load them into main memory
- Prediction
  - find the k nearest points
  - output majority class in those points
- Requires
  - distance function
  - a value of k

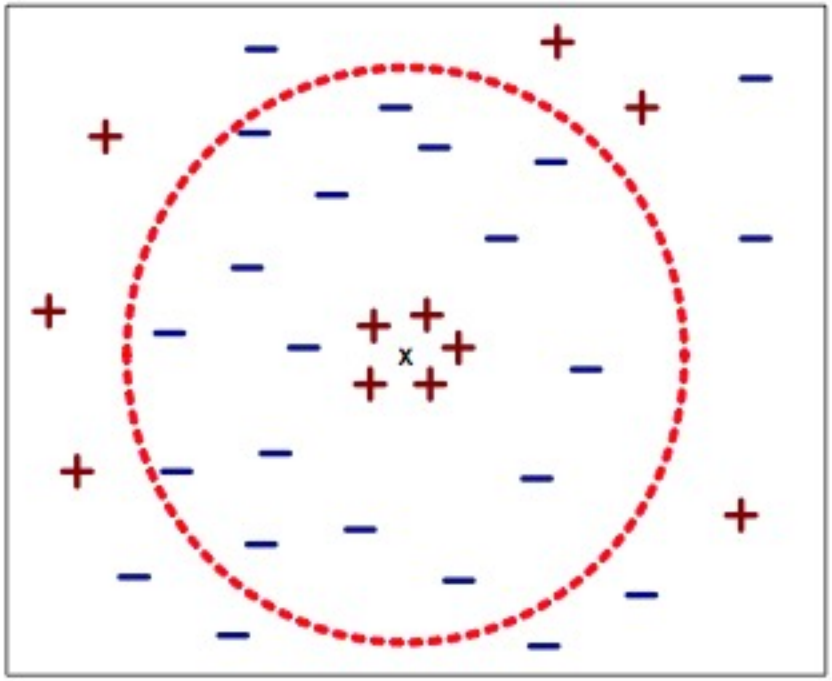


K-Nearest Neighbor Classifier

- Distance functions: Euclidean distance but also cosine similarity
- Prediction: other strategies are possible (e.g. weighted vote according to the distances).

Nearest Neighbor Classification…

- Choosing the value of k:
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other classes



Nearest Neighbor Classification…

- Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes
  - Example:
    - height of a person may vary from 1.5m to 1.8m
    - weight of a person may vary from 90lb to 300lb
    - income of a person may vary from $10K to $1M

- It suffers from the curse of dimensionality (scalability issues, data becomes too sparse)

Nearest neighbor Classification…

- k-NN classifiers are lazy learners
  - do not build models explicitly
  - Unlike eager learners such as decision tree induction and rule-based systems
  - Classifying unknown records are relatively expensive

Bayes Classifier

- A probabilistic framework for classification problems
- Conditional Probability:

$$P(C \mid A) = \frac{P(A,C)}{P(A)}$$

$$P(A \mid C) = \frac{P(A,C)}{P(C)}$$

- Bayes theorem:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Terminology in Bayesian statistics:

Posterior probability: P(C|A), P(A|C) (after A,C are taken into account)

Prior probability: P(A), P(C) (before C,A are taken into account)

Example of Bayes Theorem

- Example:
  - A doctor knows meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50000
  - Prior probability of any patient having stiff neck is 1/20

- If a patient has stiff neck, what's the probability he/she has meningitis? From Bayes it follows...

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

Bayesian Classifiers

- Attributes and class labels are random variables

- Given a record with attribute values $(a_1, a_2, \ldots, a_n)$
  - Goal is to predict the class value c_j
  - Specifically, we want to find the value c_j that maximizes $P(c_j \mid a_1, a_2, \ldots, a_n)$

- Can we estimate $P(c_j \mid a_1, a_2, \ldots, a_n)$ from data?

Bayesian Classifiers

- Approach:
  - compute the posterior probability $P(c_j \mid a_1, a_2, \ldots, a_n)$ for all values c_j using Bayes theorem

$$P(c_j|a_1, a_2, \ldots a_n) = \frac{P(a_1, a_2, \ldots, a_n|c_j)P(c_j)}{P(a_1, a_2, \ldots, a_n)}$$

  - Choose value c_j that maximizes
    $P(c_j \mid a_1, a_2, \ldots, a_n)$

  - Equivalent to choosing value of c_j that maximizes
    $P(a_1, a_2, \ldots, a_n|c_j)\, P(c_j)$

- How to estimate $P(a_1, a_2, \ldots, a_n \mid c_j)$?

Naïve Bayes Classifier

- Assume independence among a_i's when class is given:
  - $P(a_1, a_2, \ldots, a_n | c_j) = P(a_1| c_j)\, P(a_2| c_j) \ldots P(a_n| c_j)$
  - Can estimate $P(a_i| c)$ for all a_i and c_j.
  - New record is classified c_j if $P(c_j)\, \Pi\, P(a_i| c_j)$ is max.

How to Estimate Probabilities from Data?

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

- Class: $P(c_j) = N_j/N$
  - e.g., P(No) = 7/10,P(Yes) = 3/10
- For discrete attributes:

  $$P(a_i \mid c_j) = |N_{ij}|/ N_j$$

  - where $|N_{ij}|$ is number of instances having attribute a_i belonging to class c_j
  - Examples:

    P(Status=Married|No) = 4/7
    P(Refund=Yes|Yes)=0

How to Estimate Probabilities from Data?

- For continuous attributes:
  - Discretize the range into buckets
    - introduce one ordinal attribute per bucket
    - violates independence assumption
  - Two-way split: (A < v) or (A > v)
    - choose only one of the two splits as new attribute
  - Probability density estimation:
    - Assume attribute follows a normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, can use it to estimate the conditional probability $P(a_i|c_j)$

How to Estimate Probabilities from Data?

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

- Normal distribution:

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

  – One for each (A_i, c_i) pair

- For (Income, Class=No):
  – If Class=No
    - sample mean μ_{ij} = (125+100+70+120+60+220+75)/7=**110**
    - sample var. σ_{ij}^2 = **2550**

where $\sigma_{ij}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ij})^2$

Pr(Income≅120|No)=0.00775

Example of Naïve Bayes Classifier

Given a Test Record:

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

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance= 2550
If class=Yes: sample mean=90
sample variance= 16.67

- P(X|Class=No) = P(Refund=No|Class=No)
  × P(Married| Class=No)
  × P(Income=120K| Class=No)
  = 4/7 × 4/7 × 0.00775 = 0.0025

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=120K| Class=Yes)
  = 1 × 0 × ... = 0

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

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

=> Class = No

Naïve Bayes Classifier

- If one of the conditional probability is zero, then the entire expression becomes zero
- Probability estimation:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m-estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

p: prior probability

m: parameter

Example of Naïve Bayes Classifier

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

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

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

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

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

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Naïve Bayes (Summary)

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as Bayesian Belief Networks (BBN)

Platypus: is a mammal!

Mammal: (Cambridge dict.) any animal whose female feeds her young on milk from her own body.

Facts about Platypus:

- it is a mammal
- duck-billed, beaver-tailed, otter-footed
- Habitat: east Australia
- It lays eggs.
- It secretes milk from the skin (no nipples).
- hunts detecting electricity signals from the prey.
- no stomach
- male is venomous (female is not).