Naïve Bayes

A special thanks to

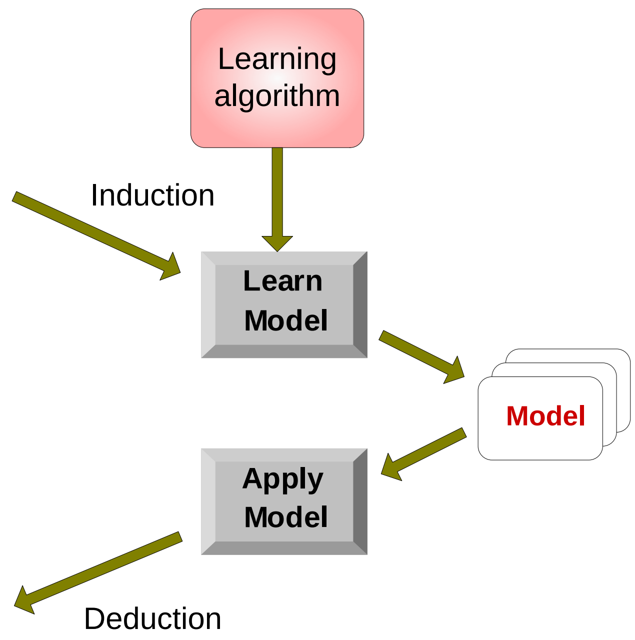
Dr. Ami Gates

Kumar et. Al 2005,2015

Reminder: A Classification Task

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

Naïve Bayes:

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

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

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

Bayes Theorem

Assume that **attributes are independent**

Let **c** be any **class** or label, Let **x** be a **data vector**

Bayes:

$P(c|\mathbf{x}) = P(\mathbf{x}|c)\ P(c)\ /\ P(\mathbf{x})$

$= P(x_1|c) * P(x_2|c) * \ldots * P(x_n|c) * P(c)\ /\ P(x_1) * P(x_2) * \ldots * P(x_n)$

Why**? Because we assume independence.**

$P(c|\mathbf{x})$ is the **posterior probability** of the class given the predictor attribute vector **x**

$P(\mathbf{x}|c)$ is the **Likelihood** – the prob of the attribute vector given the class.

$P(c)$ is the **Class Prior Probability** – the prob of the class

$P(\mathbf{x})$ is the **Predictor Prior Probability** – the prob of the attribute vector

Reminder of Conditional Prob Rules and Basic Bayes: FYI

A probabilistic framework for solving 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:

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

Example 1: Basic Bayes

Given:

A doctor knows that meningitis (M) causes stiff neck, S, is 50%. **P(S|M)=.5**

Prior probability of any patient having meningitis is 1/50000. **P(M) = 1/50000**

Prior probability of any patient having stiff neck is 1/20: **P(S) = 1/20**

Question: If a patient has stiff neck, what's the probability he/she has meningitis?

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

Bayesian Classifiers Overview

Consider each attribute and class label as random and independent variables

Given a record (data row) with attributes $(A_1, A_2, \ldots, A_n)$

- Goal is to predict class C
- Specifically, we want to find the value of C that **maximizes** $P(C| A_1, A_2, \ldots, A_n)$

Can we estimate $P(C| A_1, A_2, \ldots, A_n)$ directly from data?

Bayesian Classifiers

Approach:

Compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using Bayes

$$\boxed{P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}}$$

Choose value of C that maximizes
$P(C \mid A_1, A_2, \ldots, A_n)$

Equivalent to choosing value of C that maximizes

$P(A_1, A_2, \ldots, A_n \mid C)\ P(C)$

How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Naïve Bayes Classifier

<u>**Assume independence**</u> among attributes A_i 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 $P(A_i | C_j)$ for all A_i and C_j.

New point is classified to C_j if

$P(A_i | C_j) * P(C_j)$ is **maximal**.

Example of Naïve Bayes Classifier

Given a Test Record X, <u>classify as Evade E =Yes or No:</u>

Let record X = {Refund R=No, Married M=Yes, Income I =120K}

P(X|E=No) = P(R=No|E=No) * P(M=Yes| E=No) * P(I=120|E=No)

= 4/7 * 4/7 * 0.0072 = .0024 (see slides 12&13)

P(X|E=Yes) = P(R=No|E=Yes) * P(M=Yes| E=Yes) * P(I=120|E=Yes)

= 1 * 0 * 0.0 = 0.0

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**

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

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

→ **Class = No**

Estimating Discrete Prob 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: $P(C) = N_c/N$

$P(No) = 7/10$

$P(Yes) = 3/10$

For **discrete** attributes:

k

$P(A_i|C_k) = |A_{ik}|/ N_c$

where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k

Examples:

$P(Married{=}Yes|E{=}No) = 4/7$

$P(Refund{=}Yes|E{=}Yes) = 0$

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:

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

  – One for each (X_i, Y_i) pair

- For (Income, Class=No):
  – If Class=No
    - *sample mean = 110*
    - *sample variance = 2975*

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

The Normal Prob Dist

The Normal Prob Distribution

mu is the mean and sigma is the std dev

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

How to Estimate Probabilities from Data?

For continuous attributes:

- Discretize the range into bins
  - *one ordinal attribute per bin*
- 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* $P(A_i|c)$

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: 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

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

A: attributes

M: mammals

N: non-mammals

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

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

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

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

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)