BITS Pilani

Pilani Campus

Machine Learning

AIML CLZG565

Bayesian Learning

Raja vadhana P

Assistant Professor,

BITS - CSIS

Disclaimer and Acknowledgement

- The content for these slides has been obtained from books and various other source on the Internet
- I here by acknowledge all the contributors for their material and inputs.
- I have provided source information wherever necessary
- I have added and modified the content to suit the requirements of the course

Source: Slides of Prof. Chetana, Prof.Seetha, Prof.Sugata, Prof.Vimal, Prof.Monali, Prof. Raja vadhana , Prof.Anita from BITS Pilani , CS109 and CS229 stanford lecture notes, Tom Mitchell, Andrew Ng and many others who made their course materials freely available online

Course Plan

M1	Introduction & Mathematical Preliminaries
M2	Machine Learning Workflow
M3	Linear Models for Regression
M4	Linear Models for Classification
M5	Decision Tree
M6	Instance Based Learning
M7	Support Vector Machine
M8	Bayesian Learning
M9	Ensemble Learning
M10	Unsupervised Learning
M11	Machine Learning Model Evaluation/Comparison

Bayesian Learning
Naïve Bayes Classifier

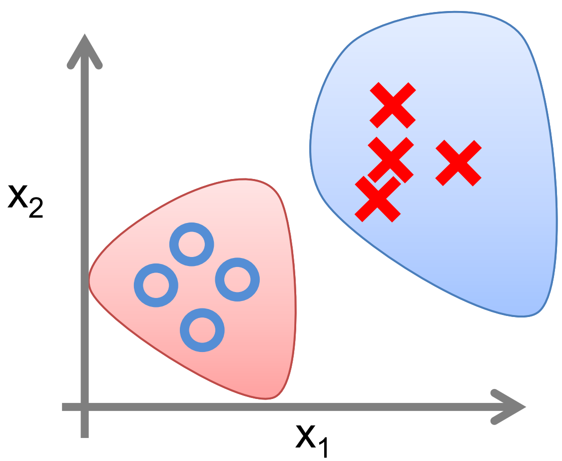
BITS Pilani

Types of Classification

Decision Theory: Interpretation

Model Building

Generative



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

Known as generative models, because by sampling from them it is possible to generate synthetic data points in the input space.

Eg., Gaussians, **Naïve Bayes**, Mixtures of multinomials , **Mixtures of Gaussians**, Bayesian networks

$$\underbrace{P(c \mid x)}_{\text{Posterior Probability}} = \frac{\overbrace{P(x \mid c)}^{\text{Likelihood}} \overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Rainy	Warm	Normal	Breeze	Same	Yes

Generative models for classification

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

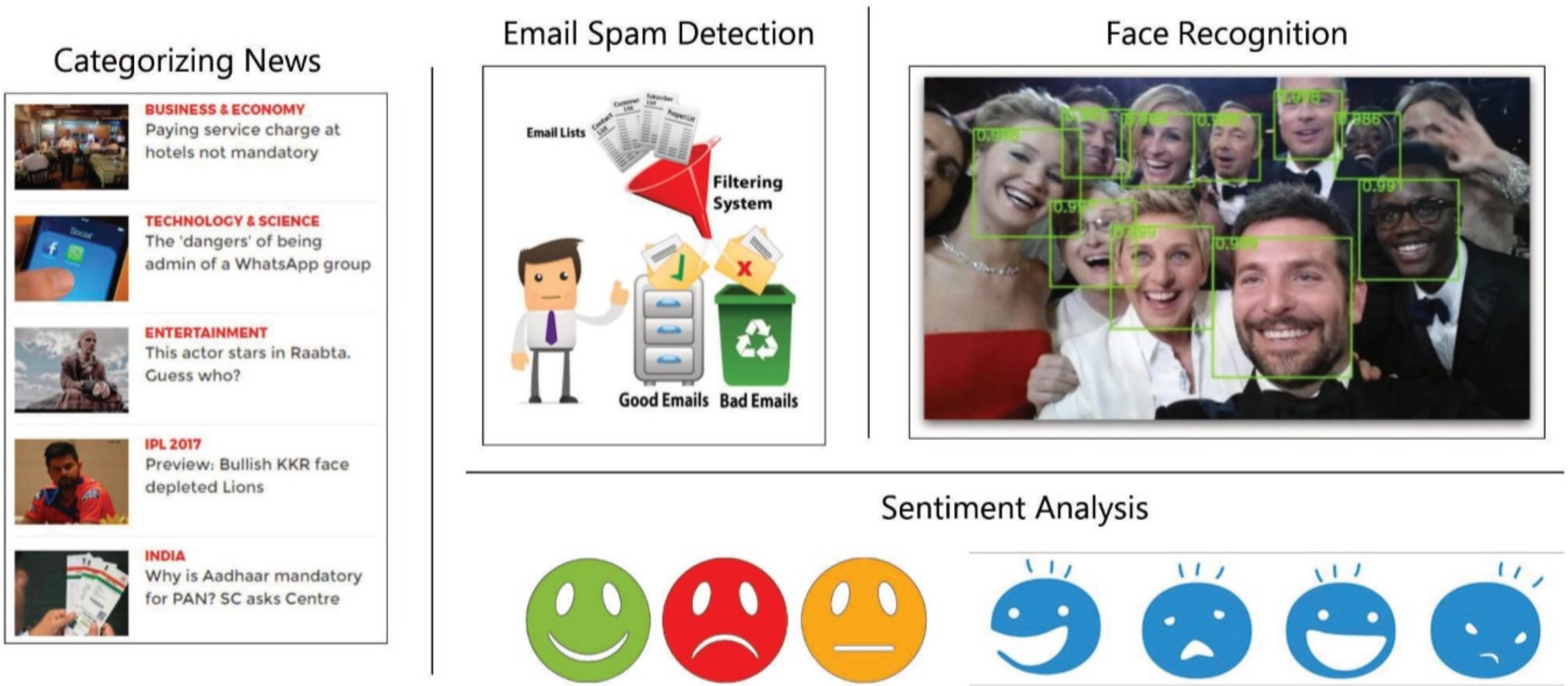
- For binary classification the denominator is given by

  p(x) = p(x|y = 1)p(y = 1) + p(x|y = 0)p(y = 0)

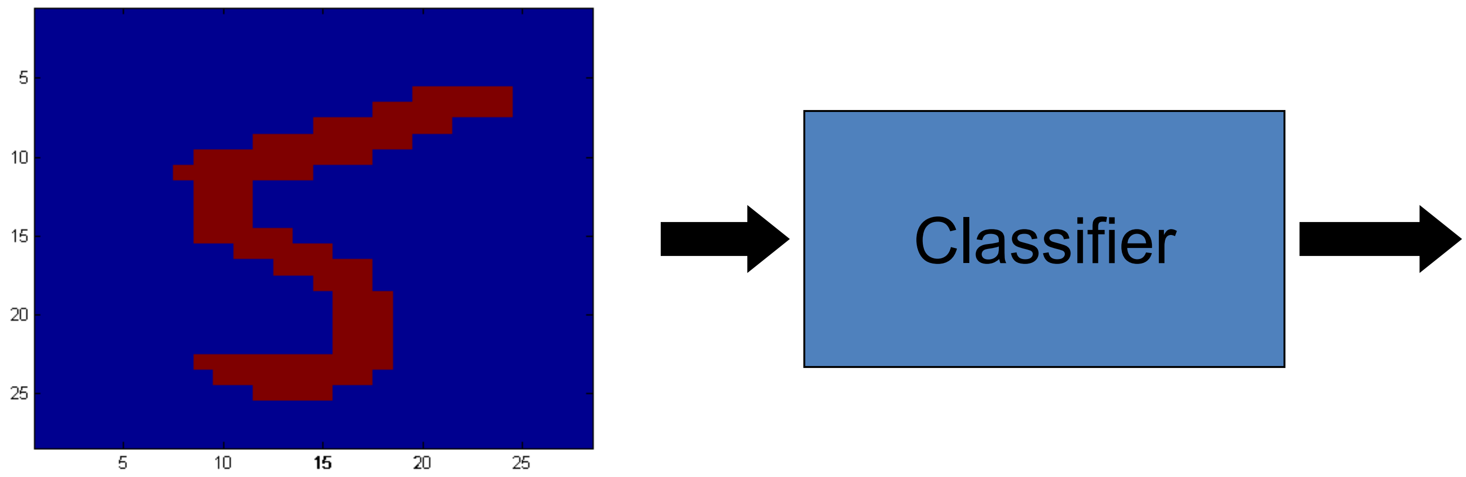
- if were calculating p(y|x) in order to make a prediction, then we don't actually need to calculate the denominator, since

$$\begin{aligned}\arg\max_y p(y|x) &= \arg\max_y \frac{p(x|y)p(y)}{p(x)} \\ &= \arg\max_y p(x|y)p(y).\end{aligned}$$

Naïve Bayes Classifier - Applications



Example: Digit Recognition



- $X_1,\ldots,X_n \in \{0,1\}$ (Black vs. White pixels)
- $Y \in \{5,6\}$ (predict whether a digit is a 5 or a 6)

The Bayes Classifier

$$
\begin{aligned}
P(Y=5|X_1,\ldots,X_n) &= \frac{P(X_1,\ldots,X_n|Y=5)P(Y=5)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)} \\
P(Y=6|X_1,\ldots,X_n) &= \frac{P(X_1,\ldots,X_n|Y=6)P(Y=6)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}
\end{aligned}
$$

- To classify, we'll simply compute these two probabilities and predict based on which one is greater

Naïve Bayes conditional Independence assumption

- Naïve Bayes assumes X_i are conditionally independent given Y

$$P(X_1|X_2,Y) = P(X_1|Y)$$

- **Assumption**:

$$P(X_1,\cdots,X_n|Y) = \prod_{j=1}^{n} P(X_j|Y)$$

i.e., X_i and X_j are conditionally independent given Y for $i \neq j$

Slide credit: Tom Mitchell

Naïve Bayes classifier: Prediction

Goal of learning P(Y|X) where X = <X_1,…, X_n>

- Bayes rule:

$$P(Y = y_k | X_1, \cdots, X_n) = \frac{P(Y = y_k) P(X_1, \cdots, X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1, \cdots, X_n | Y = y_j)}$$

- Assume conditional independence among X_i's:

$$P(Y = y_k | X_1, \cdots, X_n) = \frac{P(Y = y_k) \Pi_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \Pi_i P(X_i | Y = y_j)}$$

- Classify New Instance(x) : Pick the most probable (MAP) Y for

$$X_{new} = < X_1, \ \ldots, X_n >$$

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}} P(Y = y_k) \Pi_i P(X_i | Y = y_k)$$

Prior **Likelihood**

Slide credit: Tom Mitchell

Example: Play Tennis

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

$$\hat{Y} \leftarrow \underset{y_k}{\mathrm{argmax}}\, P(Y = y_k)\Pi_i P(X_i \mid Y = y_k)$$

$P(X|Y) \sim \mathrm{Multinom}(\pi, n) \rightarrow$ Multinomial NB (X_i – multinomial

$P(Y) \sim$ Ber(p

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Rainy	Warm	Normal	Breeze	Same	Yes

Example: Play Tennis – Learning Phase

Look up tables

Maximum likelihood estimates (MLE's):

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

Number of items in dataset D for which $Y=y_k$

Sky	Play=Yes	Play=No
Sunny	3/3	2/4
Rainy	0/3	2/4

Wind	Play=Yes	Play=No
Strong	2/3	3/4
Breeze	1/3	1/4

AirTemp	Play=Yes	Play=No
Hot	0/3	1/4
Warm	3/3	1/4
Cold	0/3	2/4

Forecast	Play=Yes	Play=No
Same	2/3	2/4
Change	1/3	2/4

Humidity	Play=Yes	Play=No
High	1/3	3/4
Normal	2/3	1/4

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Rainy	Warm	Normal	Breeze	Same	Yes

Example: Play Tennis - Testing Phase

MAP rule

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y=y_k)\prod_i P(X_i^{new}|Y=y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

P(Enjoy=Yes | X) = P(X | Enjoy=Yes). P(Enjoy=Yes) / P(X)

= P(X | Enjoy=Yes). P(Enjoy=Yes)

= P(X | Enjoy=Yes). (3/7)

= P(Sunny | Enjoy=Yes). P(Warm | Enjoy=Yes). P(Normal | Enjoy=Yes). P(Strong | Enjoy=Yes). P(Change | Enjoy=Yes).(3/7)

= (3/3) . (3/3) . (2/3) . (2/3) . (1/3) . (3/7)

=0.0635

$$\boxed{\hat{\pi}_k = \hat{P}(Y=y_k) = \frac{\#D\{Y=y_k\}}{|D|}}$$

P(Enjoy=Yes | X) > P(Enjoy=No | X) → EnjoySport = **Yes**

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Sunny	Warm	Normal	Strong	Change	????
Rainy	Warm	Normal	Breeze	Same	????

P(Enjoy=No | X) = P(X | Enjoy=No). P(Enjoy=No) / P(X)

= P(X | Enjoy=No). P(Enjoy=No)

= P(X | Enjoy=No). (4/7)

= P(Sunny | Enjoy=No). P(Warm | Enjoy=No). P(Normal | Enjoy=No). P(Strong | Enjoy=No). P(Change | Enjoy=No). (4/7)

= (2/4) . (1/4) . (1/4) . (3/4) . (2/4) . (4/7)

= 0.006696

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples) for each* value y_k

  estimate $\pi_k \equiv P(Y = y_k)$

  for each* value x_{ij} of each attribute X_i

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

Example: Play Tennis

P(Enjoy=Yes | X) = P(X | Enjoy=Yes). P(Enjoy=Yes) / P(X)

= P(X | Enjoy=Yes). P(Enjoy=Yes)

= P(X | Enjoy=Yes). (3/7)

= P(Rainy | Enjoy=Yes). P(Warm | Enjoy=Yes). P(Normal | Enjoy=Yes). P(Breeze | Enjoy=Yes). P(Same | Enjoy=Yes).(3/7)

= (0+1/3) . (3/3) . (2/3) . (1/3) . (2/3) . (3/7)

Sky	Enjoy Sport?
Sunny	Yes
Sunny	No
Rainy	No
Sunny	Yes
Sunny	No
Rainy	No
Sunny	Yes
Rainy	????

AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Warm	Normal	Strong	Same	Yes
Warm	High	Strong	Same	No
Cold	High	Strong	Change	No
Warm	Normal	Breeze	Same	Yes
Hot	Normal	Breeze	Same	No
Cold	High	Strong	Change	No
Warm	High	Strong	Change	Yes
Warm	Normal	Breeze	Same	????

P(Enjoy=No | X) = P(X | Enjoy=No). P(Enjoy=No) / P(X)

= P(X | Enjoy=No). P(Enjoy=No)

= P(X | Enjoy=No). (4/7)

= P(Rainy | Enjoy=No). P(Warm | Enjoy=No). P(Normal | Enjoy=No). P(Breeze | Enjoy=No). P(Same | Enjoy=No). (4/7)

= (2+1/4) . (1/4) . (1/4) . (1/4) . (2/4) . (4/7)

Laplace Smoothing

Smoothing

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

- Technique for smoothing categorical data.
- A small-sample correction, or **pseudo-count**, will be incorporated in every probability estimate.
- No probability will be zero.

Smoothing

Probability estimation:

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

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

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

Bayesian approach

c: number of classes

N_c: number of instances in the class

N_{ic}: number of instances having attribute value A_i in class c

p: prior probability of the class

m: constant called the **equivalent sample size**, which determines how heavily to weight p relative to the observed data

Example: Play Tennis

P(Enjoy=Yes | X) = P(X | Enjoy=Yes). P(Enjoy=Yes) / P(X)

= P(X | Enjoy=Yes). P(Enjoy=Yes)

= P(X | Enjoy=Yes). (3/7)

= P(Rainy | Enjoy=Yes). P(Warm | Enjoy=Yes). P(Normal | Enjoy=Yes). P(Breeze | Enjoy=Yes). P(Same | Enjoy=Yes).(3/7)

= (0+1/3+2) . (3/3) . (2/3) . (1/3) . (2/3) . (3/7)

$$\hat{\theta}_i = \frac{x_i + \alpha}{N + \alpha d} \qquad (i = 1, \ldots, d),$$

P(Enjoy=No | X) = P(X | Enjoy=No). P(Enjoy=No) / P(X)

= P(X | Enjoy=No). P(Enjoy=No)

= P(X | Enjoy=No). (4/7)

= P(Rainy | Enjoy=No). P(Warm | Enjoy=No). P(Normal | ... Breeze | Enjoy=No). P(Same | Enjoy=No). (4/7)

= (2+1/4+2) . (1/4) . (1/4) . (1/4) . (2/4) . (4/7)

Sky	Enjoy Sport?
Sunny	Yes
Sunny	No
Rainy	No
Sunny	Yes
Sunny	No
Rainy	No
Sunny	Yes
Rainy	????
Rainy	Yes
Sunny	Yes
Rainy	No
Sunny	No

AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Warm	Normal	Strong	Same	Yes
Warm	High	Strong	Same	No
Cold	High	Strong	Change	No
Warm	Normal	Breeze	Same	Yes
Hot	Normal	Breeze	Same	No
Cold	High	Strong	Change	No
Warm	High	Strong	Change	Yes
Warm	Normal	Breeze	Same	????

Example: Play Tennis

P(Enjoy=Yes | X) = P(X | Enjoy=Yes). P(Enjoy=Yes) / P(X)

= P(X | Enjoy=Yes). P(Enjoy=Yes)

= P(X | Enjoy=Yes). (3/7)

= P(Rainy | Enjoy=Yes). P(Warm | Enjoy=Yes). P(Normal | Enjoy=Yes). P(Breeze | Enjoy=Yes). P(Same | Enjoy=Yes).(3/7)

= (1/5) . (3/3) . (2/3) . (1/3) . (2/3) . (3/7)

=0.0127

Sky	AirTemp	Humidity	Wind	Forecast	Enjoy Sport?
Sunny	Warm	Normal	Strong	Same	Yes
Sunny	Warm	High	Strong	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	Normal	Breeze	Same	Yes
Sunny	Hot	Normal	Breeze	Same	No
Rainy	Cold	High	Strong	Change	No
Sunny	Warm	High	Strong	Change	Yes
Sunny	Warm	Normal	Strong	Change	????
Rainy	Warm	Normal	Breeze	Same	????

P(Enjoy=Yes | X) > *P(Enjoy=No | X)* → EnjoySport = **Yes**

P(Enjoy=No | X) = P(X | Enjoy=No). P(Enjoy=No) / P(X)

= P(X | Enjoy=No). P(Enjoy=No)

= P(X | Enjoy=No). (4/7)

= P(Rainy | Enjoy=No). P(Warm | Enjoy=No). P(Normal | Enjoy=No). P(Breeze | Enjoy=No). P(Same | Enjoy=No). (4/7)

= (3/6) . (1/4) . (1/4) . (1/4) . (2/4) . (4/7)

= 0.0023

Naïve Bayes: Continuous Features

- X_i can be continuous

Naïve Bayes classifier:

$$Y = \arg\max_{y} P(Y = y) \prod_{i} P(X_i | Y = y)$$

Assumption: $P(X_i | Y)$ has a **Gaussian** distribution

The Gaussian Probability Distribution

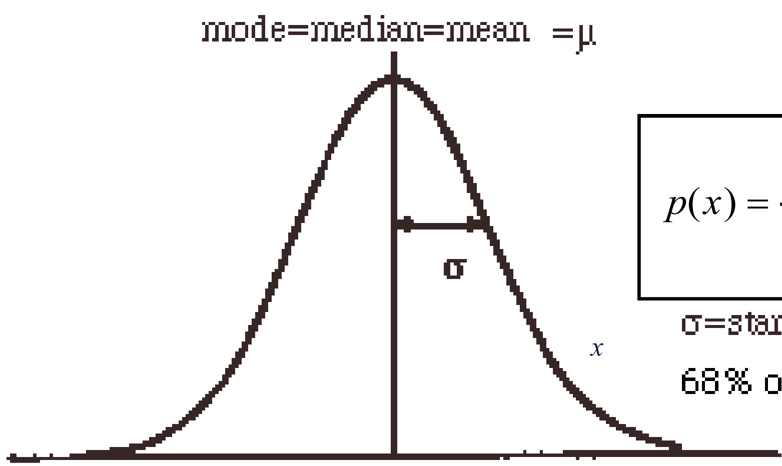
- It is a continuous distribution with pdf:

μ = mean of distribution

σ^2 = variance of distribution

x is a continuous variable $(-\infty \leq x \leq)$

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$\boxed{p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad gaussian}$$

σ=standard deviation

68% of area within ±1σ

Continuous Features : learning and prediction

innovate achieve lead

- For each target value Y_k (MLE estimate)

P$(Y = y_k) \leftarrow$ No. of instances with Y_k class/No. of Total instances

- For each attribute value X_i estimate $P\,(X_i|Y = y_k)$
  - class conditional mean , variance

- Classify New Instance(x)

Pick the most probable (MAP) Y

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}}\, P(Y = y_k)\Pi_i P(X_i|Y = y_k)$$

Continuous Features : learning

- $P(X_i|Y)$ is Gaussian
- Training: estimate mean and standard deviation
  - $\mu_i = E[X_i|Y = y]$
  - $\sigma_i^2 = E[(X_i - \mu_i)^2|Y = y]$

X_1	X_2	X_3	Y
2	3	1	1
-1.2	2	0.4	1
1.2	0.3	0	0
2.2	1.1	0	1

Continuous Features : learning

X_1	X_2	X_3	Y
2	3	1	1
-1.2	2	0.4	1
1.2	0.3	0	0
2.2	1.1	0	1

- $\mu_i = E[X_i|Y = y]$
- $\sigma_i^2 = E[(X_i - \mu_i)^2|Y = y]$

- $\mu_1 = E[X_1|Y = 1] = \frac{2+(-1.2)+2.2}{3} = 1$
- $\sigma_1^2 = E[(X_1 - \mu_1)|Y = 1] =$

$$\frac{(2-1)^2 + (-1.2\ - 1)^2 + (2.2\ - 1)^2}{3} = 2.43$$

Example: Evade Tax

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
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

P(Refund = Yes | No) = 2/6
P(Refund = No | No) = 4/6
→ P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/6
→ P(Marital Status = Divorced | No) = 0
P(Marital Status = Married | No) = 4/6
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0/3
For Taxable Income:
If class = No: sample mean = 91
sample variance = 685
If class = No: sample mean = 90
sample variance = 25

Given X = (Refund = Yes, Divorced, 120K)

P(X | No) = 2/6 X 0 X 0.0083 = 0

P(X | Yes) = 0 X 1/3 X 1.2 X 10^{-9} = 0

Naïve Bayes will not be able to classify X as Yes or No!

Example: Play Tennis

$P(X|Y) \sim N(\mu, \sigma^2) \rightarrow$ GaussianNB (X_i – real valued)

P(Enjoy=Yes | X) = P(X | Enjoy=Yes). P(Enjoy=Yes) / P(X)

= P(X | Enjoy=Yes). P(Enjoy=Yes)

= P(X | Enjoy=Yes). (3/7)

= P(Rainy | Enjoy=Yes). P(Warm | Enjoy=Yes). P(60 | Enjoy=Yes). P(Breeze | Enjoy=Yes). P(Same | Enjoy=Yes).(3/7)

= (1/3) . (3/3) . $0.15*10^{-95}$ *. (1/3) . (2/3) . (3/7)*

$$\mu_i = E[X_i|Y = yes] = 84.33$$
$$\sigma_i^2 = E[(X_i - \mu_i)^2|Y = yes] = 1.15$$

$$\mu_i = E[X_i|Y = no] = 72.5$$
$$\sigma_i^2 = E[(X_i - \mu_i)^2|Y = no] = 17.08$$

Humidity	Enjoy Sport?
85	Yes
80	No
70	No
83	Yes
90	No
50	No
85	Yes
60	????

AirTemp	Sky	Wind	Forecast	Enjoy Sport?
Warm	Sunny	Strong	Same	Yes
Warm	Sunny	Strong	Same	No
Cold	Rainy	Strong	Change	No
Warm	Rainy	Breeze	Same	Yes
Hot	Sunny	Breeze	Same	No
Cold	Rainy	Strong	Change	No
Warm	Sunny	Strong	Change	Yes
Warm	Rainy	Breeze	Same	????

P(Enjoy=No | X) = P(X | Enjoy=No). P(Enjoy=No) / P(X)

= P(X | Enjoy=No). P(Enjoy=No)

= P(X | Enjoy=No). (4/7)

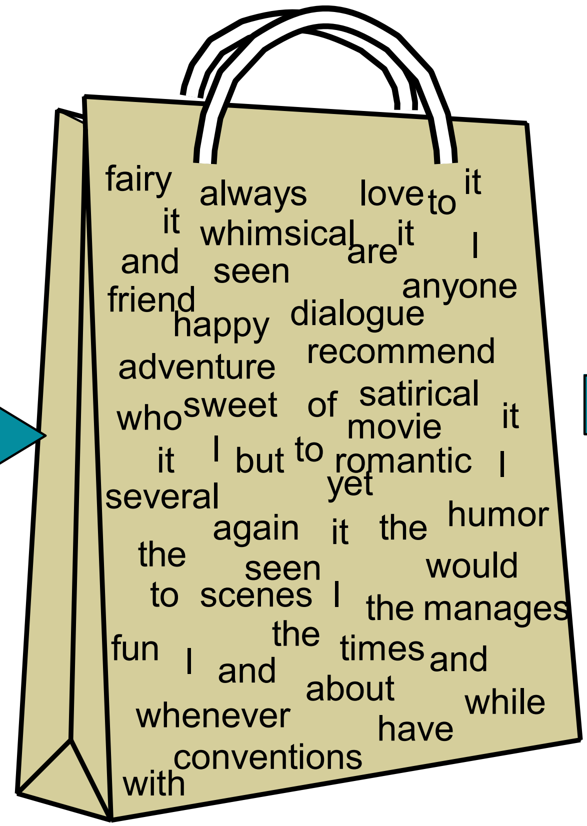
= P(Rainy | Enjoy=No). P(Warm | Enjoy=No). P(60 | Enjoy=No). P(Breeze | Enjoy=No). P(Same | Enjoy=No). (4/7)

= (2/4) . (1/4) . 0.02 . (1/4) . (2/4) . (4/7)

Text Classification using Naive Bayes Classifier

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1

Example : Multinomial model :

Which Tag sentence " A very close game" belong to?

P(Sports=Yes) = 3/5
P(Sports=No) = 2/5
P(A | Sports=Yes) = 2/11
P(A | Sports=No) = 1/9

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

P(Sports=Yes | X) = P(X | Sports=Yes). P(Sports=Yes) / P(X)
= P(X | Sports=Yes). (3/5)
= P(A| Sports=Yes). P(Very | Sports=Yes). P(Close | Sports=Yes). P(Game | Sports=Yes). (3/5)
= (2/11) . (1/11) . (0/11) . (2/11) . (3/5)

A	Great	Game	The	Election	Was	Over	Very	Clean	Match	But	Forgettable	It	Close	*Sports or Not Sports*
1	1	1												1
			1	1	1	1								0
							1	1	1					1
1		1						1		1	1			1
1				1	1							1	1	0
1		1					1						1	????

Laplace Smoothing

- Laplace smoothing: we add 1 or in general constant k to every count so it's never zero.
- To balance this, we add the number of possible words to the divisor, so the division will never be greater than 1
- In our case, the possible words are ['a', 'great', 'very', 'over', 'it', 'but', 'game', 'election', 'clean', 'close', 'the', 'was', 'forgettable', 'match'].

- In our example
- we add 1 to every probability, therefore the probability, such as **P(close | sports)**, will never be 0.

Example : Multinomial model :

Which Tag sentence " A very close game" belong to?

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

$P(Sports=Yes) = 3/5$
$P(Sports=No) = 2/5$
$P(A \mid Sports=Yes) = 2/11$
$P(A \mid Sports=No) = 1/9$

$P(Sports=Yes \mid X) = P(X \mid Sports=Yes). P(Sports=Yes) / P(X)$
$= P(X \mid Sports=Yes). (3/5)$
$= P(A \mid Sports=Yes). P(Very \mid Sports=Yes). P(Close \mid Sports=Yes). P(Game \mid Sports=Yes). (3/5)$
$= (2/11) . (1/11) . (0/11) . (2/11) . (3/5)$
$= (2+1/11+14) . (1+1/11+14) . (0+1/11+14) . (2+1/11+14) . (3/5)$
$= 0.00002765$

$(3+14/5+28)$
$=0.00002396$

A	Great	Game	The	Election	Was	Over	Very	Clean	Match	But	Forgettable	It	Close	*Sports or Not Sports*
1	1	1												1
			1	1	1	1								0
							1	1	1					1
1		1						1		1	1			1
1				1	1							1	1	0
1		1					1						1	????

Apply Laplace Smoothing

Word	P(word | Sports)	P(word | Not Sports)
a	2+1 / 11+14	1+1 / 9+14
very	1+1 / 11+14	0+1 / 9+14
close	0+1 / 11+14	1+1 / 9+14
game	2+1 / 11+14	0+1 / 9+14

$$P(a|Sports) \times P(very|Sports) \times P(close|Sports) \times P(game|Sports) \times P(Sports)$$
$$= 2.76 \times 10^{-5}$$
$$= 0.0000276$$

$$P(a|Not\,Sports) \times P(very|Not\,Sports) \times P(close|Not\,Sports) \times P(game|Not\,Sports) \times P(Not\,Sports)$$
$$= 0.572 \times 10^{-5}$$
$$= 0.00000572$$

Example 2: Multinomial model

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(w_t \mid C_k) = \frac{n_k(w_t)}{\sum_{s=1}^{|V|} n_k(w_s)},$$

N_{yes} (W=Chinese) = 5, N_{No} (W=Chinese) = 1,

|v| = 6 = {Chinese, Beijing, Shanghai, Macao, Tokyo Japan}
No of features (words) in Yes class = 8
No of features (words) in No class = 3

Example 2

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(w_t \mid C_k) = \frac{n_k(w_t)}{\sum_{s=1}^{|V|} n_k(w_s)},$$

Priors: $\hat{P}(c) = 3/4$ and $\hat{P}(\overline{c}) = 1/4$

$$\begin{aligned}
\hat{P}(\text{CHINESE}|c) &= (5+1)/(8+6) = 6/14 = 3/7 \\
\hat{P}(\text{TOKYO}|c) = \hat{P}(\text{JAPAN}|c) &= (0+1)/(8+6) = 1/14 \\
\hat{P}(\text{CHINESE}|\overline{c}) &= (1+1)/(3+6) = 2/9 \\
\hat{P}(\text{TOKYO}|\overline{c}) = \hat{P}(\text{JAPAN}|\overline{c}) &= (1+1)/(3+6) = 2/9
\end{aligned}$$

Example 2

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

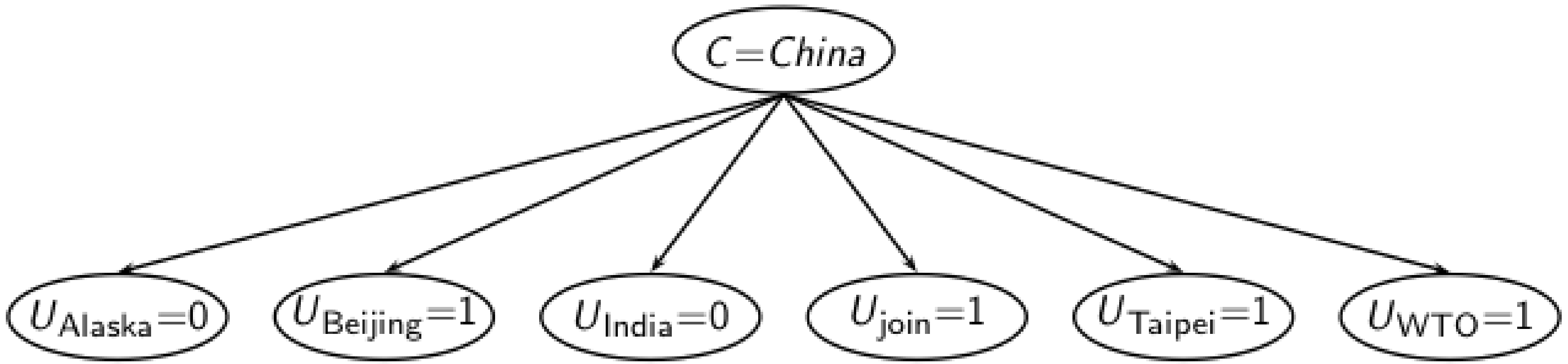
$$P(C_k|\mathcal{D}) \propto P(C_k) \prod_{j=1}^{\mathrm{len}(\mathcal{D})} P(u_j|C_k)$$

u-each word in test document

$$\hat{P}(c|d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003$$

$$\hat{P}(\overline{c}|d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001$$

Different Naive Bayes model: Bernoulli model



One feature X_w for each word in dictionary

X_w = true in document *d* if *w* appears in *d*

Example 3

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(w_t \mid C_k) = \frac{n_k(w_t)}{N_k},$$

Let $n_k(w_t)$ be the number of documents of class k in which w_t is observed; and let N_k be the total number of documents of that class.

N_{yes} (W=Chinese) = 3, N_{No} (W=Chinese) = 1,

No of features (documents) in Yes class – (N_{Yes}) = 3

No of features (documents) in No class – (N_{No}) = 1

|v| = 6

Example 3

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$
\begin{aligned}
\hat{P}(\text{Chinese}|c) &= (3+1)/(3+2) = 4/5 \\
\hat{P}(\text{Japan}|c) = \hat{P}(\text{Tokyo}|c) &= (0+1)/(3+2) = 1/5 \\
\hat{P}(\text{Beijing}|c) = \hat{P}(\text{Macao}|c) = \hat{P}(\text{Shanghai}|c) &= (1+1)/(3+2) = 2/5 \\
\hat{P}(\text{Chinese}|\overline{c}) &= (1+1)/(1+2) = 2/3 \\
\hat{P}(\text{Japan}|\overline{c}) = \hat{P}(\text{Tokyo}|\overline{c}) &= (1+1)/(1+2) = 2/3 \\
\hat{P}(\text{Beijing}|\overline{c}) = \hat{P}(\text{Macao}|\overline{c}) = \hat{P}(\text{Shanghai}|\overline{c}) &= (0+1)/(1+2) = 1/3
\end{aligned}
$$

Example 3

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

b = feature vector for the document D

b_t, = {0,1} => absence or presence of word w_t in the document

$$P(C_k \mid \mathbf{b}) \propto P(\mathbf{b} \mid C_k)\, P(C_k)$$

$$\propto P(C_k) \prod_{t=1}^{|V|} [b_t\, P(w_t \mid C_k) + (1-b_t)(1-P(w_t \mid C_k))] .$$

$$\begin{aligned} \hat{P}(c|d_5) &\propto \hat{P}(c) \cdot \hat{P}(\text{Chinese}|c) \cdot \hat{P}(\text{Japan}|c) \cdot \hat{P}(\text{Tokyo}|c) \\ &\quad \cdot (1 - \hat{P}(\text{Beijing}|c)) \cdot (1 - \hat{P}(\text{Shanghai}|c)) \cdot (1 - \hat{P}(\text{Macao}|c)) \\ &= 3/4 \cdot 4/5 \cdot 1/5 \cdot 1/5 \cdot (1-2/5) \cdot (1-2/5) \cdot (1-2/5) \\ &\approx 0.005 \end{aligned}$$

$$\begin{aligned} \hat{P}(\overline{c}|d_5) &\propto 1/4 \cdot 2/3 \cdot 2/3 \cdot 2/3 \cdot (1-1/3) \cdot (1-1/3) \cdot (1-1/3) \\ &\approx 0.022 \end{aligned}$$

Naïve Bayes classifier: Summary Model

<u>**Model:**</u> joint probability distribution given by

- $P(X, Y) = P(Y)\ P(X|Y)$
- $P(X = X_1, \cdots, X_n, Y = y_k) = P(Y = y_k)\ P(X = X_1, \cdots, X_n | Y = y_k)$

<u>**Learning/Training:**</u>

For output variable Y

- P(Y) ~ Ber(p)

For each attribute X

- P(X|Y)~ $\mathrm{Ber}(\pi)$ → Multivariate Bernoulli NB (X_i – binary)
- P(X|Y)~ $\mathrm{Multinom}(\pi, n)$ → Multinomial NB (X_i – multinomial)
- P(X|Y) ~ N(μ, σ^2) → GaussianNB (X_i – real valued)

innovate achieve lead

Logistic Regression
&
Naïve Bayes

Logistic Regression vs Naïve Bayes

Idea:

- Naïve Bayes allows computing P(Y|X) by learning P(Y) and P(X|Y)

- Why not learn P(Y|X) directly?

Logistic Regression and Gaussian Naïve Bayes Classifier

- Interestingly, the parametric form of P(Y|X) used by Logistic Regression is precisely the form implied by the assumptions of a Gaussian Naive Bayes classifier.

- Therefore, we can view Logistic Regression as a closely related alternative to GNB, though the two can produce different results in many cases

- Reference of derivation can be found in Tom Mitchell book

Where does the form come from?

- Logistic regression hypothesis representation

$$h_\theta(x) = \frac{1}{1+e^{-\theta^\top x}} = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n)}}$$

- Consider learning f: $X \rightarrow Y$, where
  - X is a vector of real-valued features $[X_1, \cdots, X_n]^\top$
  - Y is Boolean
  - Assume all X_i are conditionally independent given Y
  - Model $P(X_i|Y = y_k)$ as Gaussian $N(\mu_{ik}, \sigma_i)$
  - Model $P(Y)$ as Bernoulli π

$\boxed{\text{What is } P(Y|X_1, X_2, \cdots, X_n)\text{?}}$

Slide credit: Tom Mitchell

Where does the **form** come from?

- $$P(Y=1|X) = \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1) + P(Y=0)P(X|Y=0)}$$ Applying Bayes rule

$$= \frac{1}{1+\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}}$$ Divide by $P(Y=1)P(X|Y=1)$

$$= \frac{1}{1+\exp(\ln(\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}))}$$ Apply $\exp(\ln(\cdot))$

$$= \frac{1}{1+\exp(\ln\left(\frac{1-\pi}{\pi}\right)+\sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)})}$$ Plug in $P(X_i|Y)$

$$P(x|y_k) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_i^2}}$$

$$\sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2} X_i + \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)$$

$$P(Y=1|X_1, X_2, \cdots, X_n) = \frac{1}{1+\exp(\theta_0 + \sum_i \theta_i X_i)}$$

Slide credit: Tom Mitchell

Where does the **hypothesis function** come from?

- Logistic regression hypothesis representation

$$P(Y=1|X) = h_\theta(x) = \frac{1}{1+e^{-\theta^\top x}} = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_n x_n)}}$$

  - Model **likelihood** $\boldsymbol{P(X_i|Y=y_k)}$ as Gaussian $N(\mu_{ik}, \sigma_i)$ and assume variance is independent of class, i.e. $\sigma_{i0} = \sigma_{i1} = \sigma_i$

$$\boxed{P(x|y_k) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_i^2}}}$$

  - Model **prior** $\boldsymbol{P(Y)}$ as Bernoulli π : **P(Y=1) = π and P(Y=0) = 1-π**

What is $P(Y|X_1, X_2, \cdots, X_n)$?

Logistic Regression –Bayesian Analysis

$$P(Y=1|X)=\frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1)+P(Y=0)P(X|Y=0)}$$

Applying Bayes rule

$$P(Y=1|X)=\frac{1}{1+\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}}$$

Divide by $P(Y=1)P(X|Y=1)$

$$P(Y=1|X)=\frac{1}{1+\exp(\ln\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)})}$$

Apply $\exp(\ln(\cdot))$

$$\mathrm{P}(\mathrm{Y}=1|\mathrm{X}) = \frac{1}{1+\exp\left(\ln\frac{\mathrm{P}(\mathrm{Y}=0)}{\mathrm{P}(\mathrm{Y}=1)}+\ln\frac{\mathrm{P}(\mathrm{X}|\mathrm{Y}=0)}{\mathrm{P}(\mathrm{X}|\mathrm{Y}=1)}\right)}$$

Logistic Regression –Bayesian Analysis

By independence assumption:

P(Y=1)=π and P(Y=0)=1-π
by modelling P(Y) as Bernoulli

$$\frac{P(X|Y=0)}{P(X|Y=1)} = \prod_i \frac{P(X_i|Y=0)}{P(X_i|Y=1)}$$

$$P(Y=1|X) = \frac{1}{1 + \exp\left(\ln\frac{1-\pi}{\pi} + \ln\prod_i \frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}$$

$$P(Y=1|X) = \frac{1}{1 + \exp\left(\ln\frac{1-\pi}{\pi} + \sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}$$

Logistic Regression –Bayesian Analysis

Plug in $P(X_i|Y)$

$$P(Y=1|X) = \frac{1}{1+\exp(\ln\frac{1-\pi}{\pi} + \sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}X_i + \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right))}$$

$$P(Y=1|X) = \frac{1}{1+\exp(w_0+\sum_{i=1}^n w_i X_i)}$$

$$w_0 = \ln\frac{1-\pi}{\pi} + \sum_i \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}$$

$$w_i = \frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}$$

$$\begin{aligned}
\sum_i \ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)} &= \sum_i \ln\frac{\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(\frac{-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(\frac{-(X_i-\mu_{i1})^2}{2\sigma_i^2}\right)} \\
&= \sum_i \ln\exp\left(\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right) \\
&= \sum_i \left(\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right) \\
&= \sum_i \left(\frac{(X_i^2-2X_i\mu_{i1}+\mu_{i1}^2)-(X_i^2-2X_i\mu_{i0}+\mu_{i0}^2)}{2\sigma_i^2}\right) \\
&= \sum_i \left(\frac{2X_i(\mu_{i0}-\mu_{i1})+\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right) \\
&= \sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}X_i + \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)
\end{aligned}$$

Features of Bayesian learning

- Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct.

- Flexible approach to learning than algorithms that completely eliminate a hypothesis if it is found to be inconsistent with any single example.

Practical Issues of Bayesian learning

- Require initial knowledge of many probabilities
  - Often estimated based on background knowledge, previously available data, and assumptions about the form of the underlying distributions.
- Significant computational cost required to determine the Bayes optimal hypothesis in the general case (linear in the number of candidate hypotheses)

References

- https://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learn07-notes-nup.pdf
- https://cs229.stanford.edu/summer2019/cs229-notes2.pdf
- Tom Mitchell – Chapter 6
- https://nlp.stanford.edu/IR-book/pdf/13bayes.pdf

Thank you !

Required Reading for completed session :

T1 - Chapter # 6 (Tom M. Mitchell, Machine Learning)

R1 – Chapter # 3 (Christopher M. Bhisop, Pattern Recognition & Machine Learning)

& Refresh your MFDS & ISM parallel course basics

Next Session Plan :

Ensemble Learning

Practice Questions

Naïve Bayes – Example 1

As a part of efforts to improve students' performance in the exams, you have been given the data showing number of study hours spent by students, their gender and their final results as pass or fail. Using this sample dataset, apply Naïve Bayes classification technique, to classify the test case:

{No of study hours = 3.5, Gender="male"} either as "Pass", or "Fail".

No of study hours	Gender	Final result
4.5	Male	Pass
7	Female	Pass
2	Male	Fail
4	Female	Fail
2.5	Male	Fail
3	Female	Fail
8.3	Male	Fail
8	Female	Pass
9	Male	Pass

Naïve Bayes – Example 2

•Consider a result prediction system where student's efforts are encoded as percent of
 time a student has spent studying out of total available time.
•The input X is having just one feature representing the student's efforts having only four discrete values (25%, 50%, 75%, and 100%)
•The output Y is having 3 classes (First class, Second class, Fail)
•The priors for each class are: P(Y = First Class) = 0.5, P(Y = Second class) = 0.3, and P(Y = Fail) = 0.2.
•Based on the past data, the estimated the class-conditional probability P(X| Y) are shown in the following table.
Consider a following loss function

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}}\, l(y,\hat{y})\, P(Y = y_k) \prod_i P(X_i \mid Y = y_k)$$

Student's efforts	p(x|y=fail)	p(x|y=second class)	p(x|y=first class)
25	0.7	0.4	0.1
50	0.2	0.3	0.1
75	0.1	0.2	0.3
100	0	0.1	0.7