Naïve Bayes Classifier

- *A Naive Bayes classifier is a probabilistic machine learning model used for classification task.*
- *The root of this classifier is based on the Bayes theorem.*

Applications

- **Real time classification** — because the Naive Bayes Classifier works is very fast as compared to other classification models.
- It is used in applications that require very fast classification responses on small to medium sized datasets.
  - Spam filtering
  - Text classification
  - The Naive Bayes Classifier generally works very well with multi-class classification and even it uses that very naive assumption, it still outperforms other methods.

Outline

- Background
- Probability Basics
- Probabilistic Classification
- Naïve Bayes
  - Principle and Algorithms
  - Example: Play Tennis
- Zero Conditional Probability
- Summary

Background

- There are three methods to establish a classifier
  - a) Model a classification rule directly

    Examples: k-NN, decision trees, perceptron, SVM
  - b) Model the probability of class memberships given input data

    Example: perceptron with the cross-entropy cost
  - c) Make a probabilistic model of data within each class

    Examples: naive Bayes, model based classifiers
- *a*) and *b*) are examples of discriminative classification
- *c*) is an example of generative classification
- *b*) and *c*) are both examples of probabilistic classification

Probability Basics

- Prior, conditional and joint probability for random variables
  - Prior probability: $P(x)$
  - Conditional probability: $P(x_1 \mid x_2), P(x_2 \mid x_1)$
  - Joint probability: $\mathbf{x} = (x_1, x_2), P(\mathbf{x}) = P(x_1, x_2)$
  - Relationship: $P(x_1, x_2) = P(x_2 \mid x_1)P(x_1) = P(x_1 \mid x_2)P(x_2)$
  - Independence:

$$\boxed{P(x_2 \mid x_1) = P(x_2), P(x_1 \mid x_2) = P(x_1), P(x_1, x_2) = P(x_1)P(x_2)}$$

- Bayesian Rule

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

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Discriminative

Generative

Probability Basics

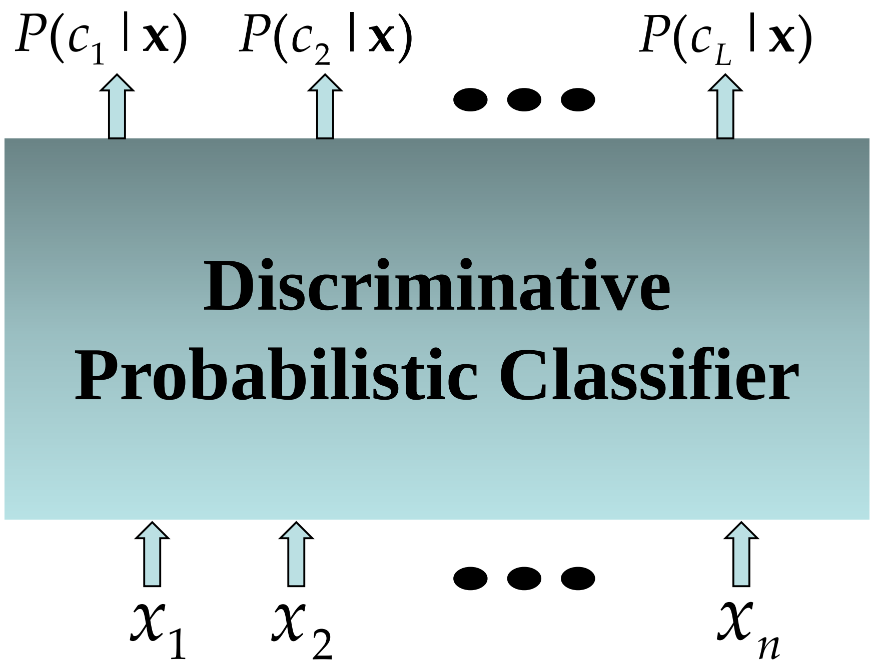
- Quiz: We have two six-sided dice. When they are tolled, it could end up with the following occurance: (A) dice 1 lands on side "3", (B) dice 2 lands on side "1", and (C) Two dice sum to eight. Answer the following questions:

1) $P(A) = ?$
2) $P(B) = ?$
3) $P(C) = ?$
4) $P(A \mid B) = ?$
5) $P(C \mid A) = ?$
6) $P(A, B) = ?$
7) $P(A, C) = ?$
8) Is $P(A, C)$ equal to $P(A) * P(C)$?

Probabilistic Classification

- Establishing a probabilistic model for classification
  - **Discriminative model**

$$P(c \mid \mathbf{x}) \quad c = c_1, \cdots, c_L, \mathbf{x} = (x_1, \cdots, x_n)$$



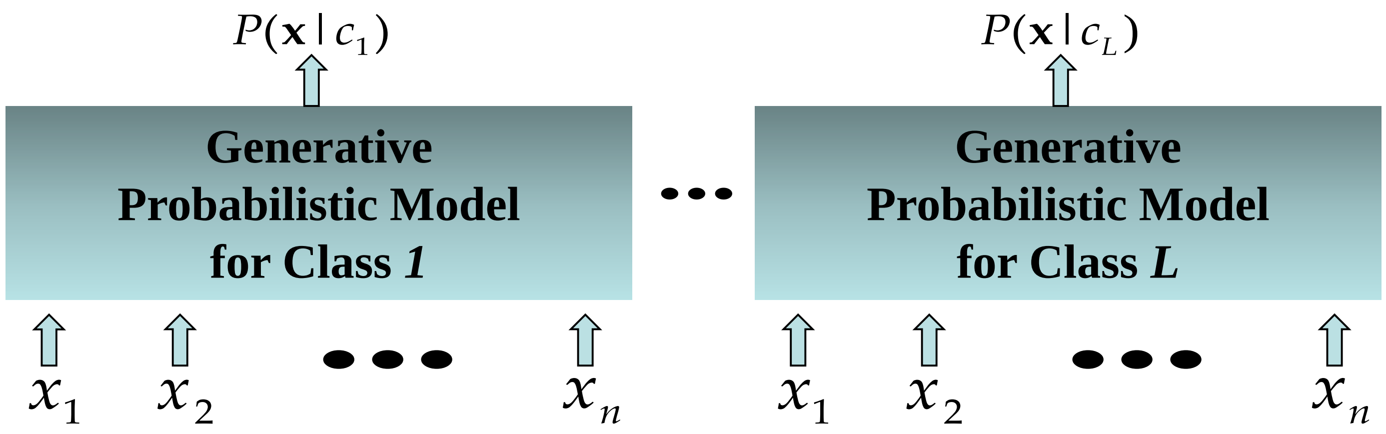
$$\mathbf{x} = (x_1, x_2, \cdots, x_n)$$

- To train a discriminative classifier regardless its probabilistic or non-probabilistic nature, all training examples of different classes must be jointly used to build up a single discriminative classifier.
- Output L probabilities for L class labels in a probabilistic classifier while a single label is achieved by a non-probabilistic classifier .

Probabilistic Classification…

- Establishing a probabilistic model for classification (cont.)
  - **Generative model (must be probabilistic)**

$$P(\mathbf{x}|c) \quad c = c_1, \cdots, c_L, \mathbf{x} = (x_1, \cdots, x_n)$$



$$\mathbf{x} = (x_1, x_2, \cdots, x_n)$$

- *L* probabilistic models have to be trained independently
- Each is trained on only the examples of the same label
- Output *L* probabilities for a given input with *L* models
- "Generative" means that such a model produces data subject to the distribution via sampling.

Probabilistic Classification…

- **M**aximum **A P**osterior (**MAP**) classification rule
  - For an input x, find the largest one from L probabilities output by a discriminative probabilistic classifier $P(c_1 \mid \mathbf{x}), \ldots, P(c_L \mid \mathbf{x})$.
  - Assign x to label c^* if $P(c^* \mid \mathbf{x})$ is the largest.
- Generative classification with the MAP rule
  - Apply Bayesian rule to convert them into posterior probabilities

$$P(c_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c_i) P(c_i)}{P(\mathbf{x})} \propto P(\mathbf{x} \mid c_i) P(c_i)$$

$$\text{for } i = 1, 2, \cdots, L$$

Common factor for all L probabilities

  - Then apply the MAP rule to assign a label

Naïve Bayes

- Bayes classification

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

- Naïve Bayes classification
  - Assume all input features are class conditionally independent!

$$X = (x_1, x_2, x_3, \ldots\ldots, x_n)$$

$$P(y|x_1, \ldots, x_n) = \frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)}$$

$$P(y|x_1, \ldots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

$$y = argmax_y P(y)\prod_{i=1}^{n} P(x_i|y)$$

Example 1

	Gender x_1	Height x_2	Class	y
$s^{(1)}$	F	1.6 m	Short	y_1
$s^{(2)}$	M	2 m	Tall	y_3
$s^{(3)}$	F	1.9 m	Medium	y_2
$s^{(4)}$	F	1.88 m	Medium	y_2
$s^{(5)}$	F	1.7 m	Short	y_1
$s^{(6)}$	M	1.85 m	Medium	y_2
$s^{(7)}$	F	1.6 m	Short	y_1
$s^{(8)}$	M	1.7 m	Short	y_1
$s^{(9)}$	M	2.2 m	Tall	y_3
$s^{(10)}$	M	2.1 m	Tall	y_3
$s^{(11)}$	F	1.8 m	Medium	y_2
$s^{(12)}$	M	1.95 m	Medium	y_2
$s^{(13)}$	F	1.9 m	Medium	y_2
$s^{(14)}$	F	1.8 m	Medium	y_2
$s^{(15)}$	F	1.75 m	Medium	y_2

Consider the data given in table and classify a sample x={M, 1.95m)

Example 1

Table 3.2 Number of training samples, $N_{qv_{lx_j}}$, of class q having value v_{lx_j}

Value v_{lx_j}	Count $N_{qv_{lx_j}}$		
	Short $q = 1$	Medium $q = 2$	Tall $q = 3$
v_{1x_1}: M	1	2	3
v_{2x_1}: F	3	6	0
v_{1x_2}: (0, 1.6] bin	2	0	0
v_{2x_2}: (1.6, 1.7] bin	2	0	0
v_{3x_2}: (1.7, 1.8] bin	0	3	0
v_{4x_2}: (1.8, 1.9] bin	0	4	0
v_{5x_2}: (1.9, 2.0] bin	0	1	1
v_{6x_2}: (2.0, ∞] bin	0	0	2

Consider the data given in table and classify a sample x={M, 1.95m)

Example 1…

$$P(y_1) = \frac{4}{15} \qquad P(y_2) = \frac{8}{15} \qquad P(y_3) = \frac{3}{15}$$

$$P(x_1/y_1) = \frac{1}{4}, P(x_1/y_2) = \frac{2}{8}, P(x_1/y_3) = 3/3$$

$$P(x_2/y_1) = \frac{0}{4}, P(x_2/y_2) = \frac{1}{8}, P(x_2/y_3) = \frac{1}{3}$$

$$P\left(\frac{x}{y_1}\right) = P\left(\frac{x_1}{y_1}\right) \times P\left(\frac{x_2}{y_1}\right) = \frac{1}{4} \times 0 = 0$$

Example 1...

$$P\left(\frac{x}{y_2}\right) = P\left(\frac{x_1}{y_2}\right) \times P\left(\frac{x_2}{y_2}\right) = \frac{2}{8} \times \frac{1}{8} = 1/32$$

$$P\left(\frac{x}{y_3}\right) = P\left(\frac{x_1}{y_3}\right) \times P\left(\frac{x_2}{y_3}\right) = \frac{3}{3} \times \frac{1}{3} = 1/3$$

$$P\left(\frac{x}{y_1}\right) \times P(y_1) = 0 \times \frac{4}{15} = 0$$

Example 1…

$$P\left(\frac{x}{y_2}\right) \times P(y_2) = \frac{1}{32} \times \frac{8}{15} = 0.0166$$

$$P\left(\frac{x}{y_3}\right) \times P(y_3) = \frac{1}{3} \times \frac{3}{15} = 0.66$$

Using MAP

$$y_{NB} = \arg Max\, P(\frac{x}{y_q}) \times P(y_q)$$

Example 2

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

X′=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)

Example 2...

- Learning Phase

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=*No*
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

$P(\text{Play}=\textit{Yes}) = 9/14 \qquad P(\text{Play}=\textit{No}) = 5/14$

$\mathbf{x}'$=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)

Example 2…

- Test Phase
  - Given a new instance, predict its label

    $\mathbf{x}'$=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)
  - Look up tables achieved in the learning phrase

P(Outlook=*Sunny*|Play=*Yes*) = 2/9	P(Outlook=*Sunny*|Play=*No*) = 3/5
P(Temperature=*Cool*|Play=*Yes*) = 3/9	P(Temperature=*Cool*|Play==*No*) = 1/5
P(Huminity=*High*|Play=*Yes*) = 3/9	P(Huminity=*High*|Play=*No*) = 4/5
P(Wind=*Strong*|Play=*Yes*) = 3/9	P(Wind=*Strong*|Play=*No*) = 3/5
P(Play=*Yes*) = 9/14	P(Play=*No*) = 5/14

  - Decision making with the MAP rule

$$P(Yes|\mathbf{x}') \approx [P(Sunny|Yes)P(Cool|Yes)P(High|Yes)P(Strong|Yes)]P(Play{=}Yes) = 0.0053$$

$$P(No|\mathbf{x}') \approx [P(Sunny|No)\ P(Cool|No)P(High|No)P(Strong|No)]P(Play{=}No) = 0.0206$$

Given the fact $P(Yes|\mathbf{x}') < P(No|\mathbf{x}')$, we label $\mathbf{x}'$ to be "*No*".

Zero Conditional Probability…

- If no example contains the feature value
  - In this circumstance, we face a zero conditional probability problem during test

$$\hat{P}(x_1 | c_i) \cdots \hat{P}(a_{jk} | c_i) \cdots \hat{P}(x_n | c_i) = 0 \quad \text{for } x_j = a_{jk},\ \hat{P}(a_{jk} | c_i) = 0$$

  - For a remedy, class conditional probabilities re-estimated with

$$\hat{P}(a_{jk} | c_i) = \frac{n_c + mp}{n + m}$$

(m-estimate)

n_c : number of training examples for which $x_j = a_{jk}$ and $c = c_i$

n : number of training examples for which $c = c_i$

p : prior estimate (usually, $p = 1/t$ for t possible values of x_j)

m : weight to prior (number of "virtual" examples, $m \geq 1$)

Zero conditional probability…

- Example: P(outlook=overcast|no)=0 in the play-tennis dataset
  - Adding *m* "virtual" examples (*m*: up to 1% of #training example)
    - In this dataset, # of training examples for the "no" class is 5.
    - We can only add ***m*=1** "virtual" example in our m-esitmate remedy.
  - The "outlook" feature can takes only 3 values. So *p*=1/3.
  - Re-estimate P(outlook|no) with the m-estimate

$$P(\text{overcast}|\text{no}) = \frac{0+1*\left(\frac{1}{3}\right)}{5+1} = \frac{1}{18}$$

$$P(\text{sunny}|\text{no}) = \frac{3+1*\left(\frac{1}{3}\right)}{5+1} = \frac{5}{9} \qquad P(\text{rain}|\text{no}) = \frac{2+1*\left(\frac{1}{3}\right)}{5+1} = \frac{7}{18}$$

Summary

- Naïve Bayes: the conditional independence assumption
  - Training and test are very efficient
  - Two different data types lead to two different learning algorithms
  - Working well sometimes for data violating the assumption!
- A popular generative model
  - Performance competitive to most of state-of-the-art classifiers even in presence of violating independence assumption
  - Many successful applications, e.g., spam mail filtering
  - A good candidate of a base learner in ensemble learning
  - Apart from classification, naïve Bayes can do more…