

# Machine learning

Prepared by : Dr. Hanaa Bayomi  
Updated By: Prof Abeer ElKorany



## Lecture 5: Naïve bayse

# Classification Paradigms

---

- 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: Regression model
  - 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

# Generative vs Discriminative Models

---

## Notation:





- **Generative Model:** A model that defines joint features and class, in other words algorithms that try to define what kind of data we expect to see in each class.
- probability distribution, For example, Naïve Bayes.

**Discriminative Model:** A model that defines class conditional probability distribution, they try to learn mappings directly from the space of inputs  $X$  to the labels  $\{0, 1\}$  (Try to learn  $p(y|x)$  )

- For example, Logistic Regression.

# Generative vs Discriminative Models

---

	Generative Models (ex. Naïve Bayes)	Discriminative Models (ex. Logistic Regression)
Classification Accuracy		
Missing Features		

# Naïve bayes classifier

---

- It is a classification technique based on *Bayes theorem* with *independent assumption among features (predictors)*.
- Naïve Bayes model is easy to build, with no complicated iterative parameter estimation *which makes it particularly useful for very large datasets*

# Probability Basics

---

- Prior, conditional and joint probability
  - Prior probability:  $P(X)$
  - Conditional probability:  $P(X_1 | X_2), P(X_2 | X_1)$
  - Joint probability:  $\mathbf{X} = (X_1, X_2), P(\mathbf{X}) = P(X_1, X_2)$
  - Relationship:  $P(X_1, X_2) = P(X_2 | X_1)P(X_1) = P(X_1 | X_2)P(X_2)$
  - Independence:  $P(X_2 | X_1) = P(X_2), P(X_1 | X_2) = P(X_1), P(X_1, X_2) = P(X_1)P(X_2)$
- Bayesian Rule

$$P(C | \mathbf{X}) = \frac{P(\mathbf{X} | C)P(C)}{P(\mathbf{X})} \quad \text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

# Bayes Theorem

- Given a class  $C$  and feature  $X$  which bears on the class:

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

- $P(C)$  : independent probability of  $C$  (*hypotheses*): *prior probability*
- $P(X)$  : independent probability of  $X$  (*data, predictor*)
- $P(X/C)$ : conditional probability of  $X$  given  $C$ : *likelihood*
- $P(C/X)$ : conditional probability of  $C$  given  $X$ : *posterior probability*

# Maximum A Posterior

- Based on Bayes Theorem, we can compute the *Maximum A Posterior* (MAP) hypothesis for the data
- We are interested in the best hypothesis for some space  $C$  given observed training data  $X$ .

$$\begin{aligned}c_{MAP} &\equiv \operatorname{argmax}_{c \in C} P(c | X) \\&= \operatorname{argmax}_{c \in C} \frac{P(X | c)P(c)}{P(X)} \\&= \operatorname{argmax}_{c \in C} P(X | c)P(c)\end{aligned}$$

$C$ : set of all hypothesis (Classes).

Note that we can drop  $P(X)$  as the probability of the data is constant (and independent of the hypothesis).



# Bayes Classifiers

**Assumption:** training set consists of instances of different classes described  $c_j$  as conjunctions of attributes values

**Task:** Classify a new instance  $d$  based on a tuple of attribute values into one of the classes  $c_j \in \mathcal{C}$

**Key idea:** assign the most probable class  $c_{MAP}$  using Bayes Theorem.

$$\begin{aligned} c_{MAP} &= \operatorname{argmax}_{c_j \in \mathcal{C}} P(c_j \mid x_1, x_2, \dots, x_n) \\ &= \operatorname{argmax}_{c_j \in \mathcal{C}} \frac{P(x_1, x_2, \dots, x_n \mid c_j) P(c_j)}{P(x_1, x_2, \dots, x_n)} \\ &= \operatorname{argmax}_{c_j \in \mathcal{C}} P(x_1, x_2, \dots, x_n \mid c_j) P(c_j) \end{aligned}$$

# The Naïve Bayes Model

---

- Naïve Bayes classification
  - Making the assumption that **all input attributes are independent**
- The *Naïve Bayes Assumption*: Assume that the effect of the value of the predictor (X) on a given class ( C ) is independent of the values of other predictors.
- This assumption is called class conditional independence

$$P(x_1, x_2, \dots, x_n \mid C) = P(x_1 \mid C) \times P(x_2 \mid C) \times \dots \times P(x_n \mid C)$$

$$P(x_1, x_2, \dots, x_n \mid C) = \prod_{i=1}^n P(x_i \mid C)$$

# Naïve Bayes Algorithm

- Naïve Bayes Algorithm (for discrete input attributes) has two phases

- **1. Learning Phase:** Given a training set  $S$ ,

Learning is easy, just create probability tables.

For each target value of  $c_i$  ( $c_i = c_1, \dots, c_L$ )

$\hat{P}(C = c_i) \leftarrow$  estimate  $P(C = c_i)$  with examples in  $S$ ;

For every attribute value  $x_{jk}$  of each attribute  $X_j$  ( $j = 1, \dots, n; k = 1, \dots, N_j$ )

$\hat{P}(X_j = x_{jk} | C = c_i) \leftarrow$  estimate  $P(X_j = x_{jk} | C = c_i)$  with examples in  $S$ ;

Output: conditional probability tables; for  $X_j, N_j \times L$  elements

- **2. Test Phase:** Given an unknown instance  $\mathbf{X}' = (a'_1, \dots, a'_n)$ ,

Look up tables to assign the label  $c^*$  to  $\mathbf{X}'$  if

$$[\hat{P}(a'_1 | c^*) \cdots \hat{P}(a'_n | c^*)] \hat{P}(c^*) > [\hat{P}(a'_1 | c) \cdots \hat{P}(a'_n | c)] \hat{P}(c), \quad c \neq c^*, c = c_1, \dots, c_L$$

Classification is easy, just multiply probabilities

# How to Estimate Probabilities from Data?

$T$	$R$	$P(T, R)$
hot	no rain	0.3
hot	rain	0.2
cold	no rain	0.3
cold	rain	0.2

=

×

$T$	$P(T)$
hot	0.5
cold	0.5

$R$	$P(R)$
no rain	0.6
rain	0.4

# How to Estimate Probabilities from Data?

<i>Tid</i>	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:  $P(C) = N_c / N$ 
  - e.g.,  $P(\text{No}) = 7/10$ ,  
 $P(\text{Yes}) = 3/10$
- For discrete Attribute/Features:
$$P(A_i \mid C_k) = |A_{ik}| / N_{c_k}$$
  - where  $|A_{ik}|$  is number of instances having attribute  $A_i$  and belongs to class  $C_k$
  - Examples:  
 $P(\text{Status}=\text{Married} \mid \text{No}) = 4/7$   
 $P(\text{Refund}=\text{Yes} \mid \text{Yes})=0$

# Example1

- Example: Play Tennis

*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

# Example1

- Learning Phase

<i>Outlook</i>	Play=Yes	Play=No
<i>Sunny</i>	2/9	3/5
<i>Overcast</i>	4/9	0/5
<i>Rain</i>	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
<i>High</i>	3/9	4/5
<i>Normal</i>	6/9	1/5

Wind	Play=Yes	Play=No
<i>Strong</i>	3/9	3/5
<i>Weak</i>	6/9	2/5

$$P(\text{Play=Yes}) = 9/14$$

$$P(\text{Play=No}) = 5/14$$

# Example1

- **Test Phase**

- Given a new instance, predict its label

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

- **Look up tables achieved in the learning phrase**

$$P(\text{Outlook}=\textit{Sunny} \mid \text{Play}=\textit{Yes}) = 2/9$$

$$P(\text{Outlook}=\textit{Sunny} \mid \text{Play}=\textit{No}) = 3/5$$

$$P(\text{Temperature}=\textit{Cool} \mid \text{Play}=\textit{Yes}) = 3/9$$

$$P(\text{Temperature}=\textit{Cool} \mid \text{Play}=\textit{No}) = 1/5$$

$$P(\text{Humidity}=\textit{High} \mid \text{Play}=\textit{Yes}) = 3/9$$

$$P(\text{Humidity}=\textit{High} \mid \text{Play}=\textit{No}) = 4/5$$

$$P(\text{Wind}=\textit{Strong} \mid \text{Play}=\textit{Yes}) = 3/9$$

$$P(\text{Wind}=\textit{Strong} \mid \text{Play}=\textit{No}) = 3/5$$

$$P(\text{Play}=\textit{Yes}) = 9/14$$

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

- **Decision making with the MAP rule**

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

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

Given the fact  $P(\textit{Yes} \mid \mathbf{x}') < P(\textit{No} \mid \mathbf{x}')$ , we label  $\mathbf{x}'$  to be “No”.



# Example2

Training set

<i>Tid</i>	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 Record:

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

## Example2 of Naïve Bayes Classifier

naive Bayes Classifier:

$$P(\text{Refund}=\text{Yes}|\text{No}) = 3/7$$

$$P(\text{Refund}=\text{No}|\text{No}) = 4/7$$

$$P(\text{Refund}=\text{Yes}|\text{Yes}) = 0$$

$$P(\text{Refund}=\text{No}|\text{Yes}) = 1$$

$$P(\text{Marital Status}=\text{Single}|\text{No}) = 2/7$$

$$P(\text{Marital Status}=\text{Divorced}|\text{No}) = 1/7$$

$$P(\text{Marital Status}=\text{Married}|\text{No}) = 4/7$$

$$P(\text{Marital Status}=\text{Single}|\text{Yes}) = 2/7$$

$$P(\text{Marital Status}=\text{Divorced}|\text{Yes}) = 1/7$$

$$P(\text{Marital Status}=\text{Married}|\text{Yes}) = 0$$

For taxable income:

If class=No:      sample mean=110  
                         sample variance=2975

If class=Yes:      sample mean=90  
                         sample variance=25

Given a Test Record:

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

$$\begin{aligned} \square \quad P(X|\text{Class}=\text{No}) &= P(\text{Refund}=\text{No}|\text{Class}=\text{No}) \\ &\quad \times P(\text{Married}|\text{Class}=\text{No}) \\ &\quad \times P(\text{Income}=120\text{K}|\text{Class}=\text{No}) \\ &= 4/7 \times 4/7 \times 0.0072 = 0.0024 \end{aligned}$$

$$\begin{aligned} \square \quad P(X|\text{Class}=\text{Yes}) &= P(\text{Refund}=\text{No}|\text{Class}=\text{Yes}) \\ &\quad \times P(\text{Married}|\text{Class}=\text{Yes}) \\ &\quad \times P(\text{Income}=120\text{K}|\text{Class}=\text{Yes}) \\ &= 1 \times 0 \times 1.2 \times 10^{-9} = 0 \end{aligned}$$

Since  $P(X|\text{No})P(\text{No}) > P(X|\text{Yes})P(\text{Yes})$

Therefore  $P(\text{No}|X) > P(\text{Yes}|X)$

$\Rightarrow \text{Class} = \text{No}$

# Continuous-valued Input Attributes

---

- Numberless values for an attribute
- Conditional probability modeled with the normal distribution

$$\hat{P}(X_j | C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2}\right)$$

$\mu_{ji}$  : mean (average) of attribute values  $X_j$  of examples for which  $C = c_i$

$\sigma_{ji}$  : standard deviation of attribute values  $X_j$  of examples for which  $C = c_i$

- Learning Phase: for  $\mathbf{X} = (X_1, \dots, X_n)$ ,  $C = c_1, \dots, c_L$   
Output:  $n \times L$  normal distributions and  $P(C = c_i) \ i = 1, \dots, L$
- Test Phase: for  $\mathbf{X}' = (X'_1, \dots, X'_n)$ 
  - Calculate conditional probabilities with all the normal distributions
  - Apply the MAP rule to make a decision

# Naïve Bayes: Continuous Features

---

- $P(X_i|Y)$  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

# Naïve Bayes: Continuous Features

- $P(X_i|Y)$  is Gaussian
- Training: estimate mean and standard deviation

- $\mu_i = E[X_i|Y = y]$

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

- $\mu_{11} = E[X_1|Y = 1] = \frac{2+(-1.2)+2.2}{3} = 1$

- $\mu_{10} = E[X_1|Y = 0] = \frac{1.2}{1} = 1.2$

- $\sigma_{11}^2 = E[(X_1 - \mu_1)|Y = 1] = \frac{(2-1)^2+(-1.2-1)^2+(2.2-1)^2}{3} = 2.43$

- $\sigma_{10}^2 = E[(X_1 - \mu_1)|Y = 0] = \frac{(1.2-1.2)^2}{1} = 0$

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

# Naïve Bayes

- Example: Continuous-valued Features

- Temperature is naturally of continuous value.

**Yes:** 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8

**No:** 27.3, 30.1, 17.4, 29.5, 15.1

- Estimate mean and variance for each class

$$\mu = \frac{1}{N} \sum_{n=1}^N x_n, \quad \sigma^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2$$

$$\mu_{Yes} = 21.64, \quad \sigma_{Yes} = 2.35$$

$$\mu_{No} = 23.88, \quad \sigma_{No} = 7.09$$

- **Learning Phase:** output two Gaussian models for  $P(\text{temp}|\text{C})$

$$\hat{P}(x | Yes) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{2 \times 2.35^2}\right) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{11.09}\right)$$

$$\hat{P}(x | No) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{2 \times 7.09^2}\right) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{50.25}\right)$$

# 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

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

$n_c$  : 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:  $P(\text{outlook}=\text{overcast}|\text{no})=0$  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.

<i>Outlook</i>	Play=Yes	Play=No
<i>Sunny</i>	2/9	3/5
<i>Overcast</i>	4/9	0/5
<i>Rain</i>	3/9	2/5

We can only add  $m=1$  “virtual” example in our m-estimate remedy.



# Zero conditional probability

- The “outlook” feature can takes only 3 values. So  $p=1/3$ .
- Re-estimate  $P(\text{outlook}|\text{no})$  with the m-estimate

$$P(\text{Outlook}=\text{Sunny}|\text{Play}=\text{No}) = 3/5$$

$$P(\text{Outlook}=\text{overcast}|\text{Play}=\text{No}) = 0$$

$$P(\text{Outlook}=\text{Rain}|\text{Play}=\text{No}) = 2/5$$

$$P(\text{overcast}|\text{no}) = \frac{0+1*\left(\frac{1}{3}\right)}{5+1} = \frac{1}{6}$$

$$P(\text{sunny}|\text{no}) = \frac{3+1*\left(\frac{1}{3}\right)}{5+1} = \frac{5}{6} \quad P(\text{rain}|\text{no}) = \frac{2+1*\left(\frac{1}{3}\right)}{5+1} = \frac{5}{6}$$

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

$n_c$  : 0 (No.of samples **outlook=overcast|no**)

$n$  : 5 (No.of samples **class=no**)

$p$  : **1/3** (outlook has 3 values(sunny, overcast, rain) )

$m$  : **1**

# Conclusion

- Naïve Bayes is based on the **independence assumption**
- **Training** is very easy and fast; just requiring considering each attribute in each class separately
- **Test** is straightforward; just looking up tables or calculating conditional probabilities with normal distributions
- Naïve Bayes is a popular generative model
- Performance of naïve Bayes is **competitive** to most of state-of-the-art classifiers even if in presence of violating the independence assumption
- It has many successful applications, e.g., spam mail filtering