# Lecture 2.4-2.5

- Naive Bayes Classifier

# Classifier Establishment

- Model a classification rule directly
  - k-NN, decision trees, perceptron, SVM
- Model the probability of class memberships given input data
  - multi-layered perceptron with the cross-entropy cost
- Make a probabilistic model of data within each class
  - Naïve Bayes

# 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

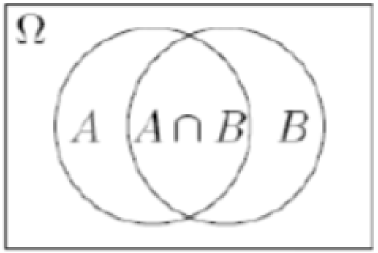
$$P(C | \mathbf{X}) = \frac{P(\mathbf{X} | C)P(C)}{P(\mathbf{X})} \quad Posterior = \frac{Likelihood \times Prior}{Evidence}$$

# Probabilities Meaning

- $P(A)$ denotes our belief that event A will happen
  - $P(A) = 0$; means that the event never occurs
  - $P(A) = 1$;means that the even always occurs
- The sum of the probabilities of all events equals to one
  - If $\overline{P(A)}$ represents the event of $P(A)$ not occuring, then

$$P(A) + \overline{P(A)} = 1$$

# Joint Probability




- Joint probability refers to the probability of **two or more events occurring simultaneously**
- For two events $A$ and $B$, the joint probability $P(A \cap B)$ is the probability that both events $A$ and $B$ happen at the same time
- If $A$ and $B$ are independent events:

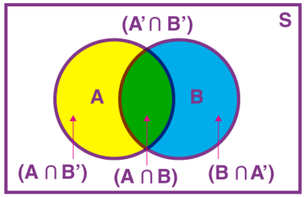
$$P(A \cap B) = P(A).P(B)$$

- If $A$ and $B$ are dependent events:

$$P(A \cap B) = P(A).P(B|A) = P(B).P(A|B)$$

Here, $P(B|A)$ is the conditional probability of $B$ given that $A$ has occurred, vice versa for $P(A \mid B)$

# The Law of Total Probability




- The law of total probability is

$$P(A) = \sum_n P(A \cap B_n) = \sum_n P(A|B_n).P(B_n)$$

$\{B_n: n = 1,2,3..\}$is a finite or countable infinite set of mutually exclusive

- Also,

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$$

# Bayes' Rule

- **Bayes' Rule**, also known as Bayes' Theorem, is a formula used to calculate the probability of an event based on prior knowledge of conditions that might affect the event

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(A \mid B)$: Posterior probability of $A$ given $B$

$P(B \mid A)$: Likelihood of $B$ given $A$

$P(A)$: Prior probability of $A$

$P(B)$: Evidence

# Example 1

A patient comes to a doctor with a symptom $S$. The doctor knows that 30% of the people with this symptom have a disease $D$ from which they will die if they are not operated on immediately. However, the operation is dangerous with 50% fatality irrespective of whether the patient has the disease or not. The doctor knows that 60% of people with $D$ have blood type $A$ while only 30% of the normal population has this blood type. On performing a blood test it is found that the patient has blood type $A$. Should the doctor operate?

# Solution 1

- Probability of having the disease given symptom $S$ : $\boldsymbol{P(D|S) = 0.30}$
- Probability of not having the disease (complement of $D$) given symptom:
$$\boldsymbol{P(D^c|S) = 1 - 0.30 = 0.70}$$
- Probability of having blood type $A$ given disease $D$
$$\boldsymbol{P(A|D) = P(A|D) = 0.60}$$
- Probability of having blood type $A$ without the disease
$$\boldsymbol{P\left(A\middle|D^C\right) = 0.30}$$
- To Calculate P(D|A)
$$\boldsymbol{P(D|A) = \frac{P(A|D).P(D)}{P(A)}}$$
- Calculate P(A):
$$\boldsymbol{P(A) = P(A|D)\cdot P(D|S) + P(A|D^c)\cdot P(D^c|S) = 0.6\times 0.3 + 0.3\times 0.7 = 0.39}$$
- $\boldsymbol{P(D|A) = \frac{P(A|D).P(D)}{P(A)} = \frac{0.6\times 0.3}{0.39} = 0.4615}$
- $\boldsymbol{P(Survival) = 1 - P(D|A) = 0.5385 > 50\%\ fatality}$

# The Naïve Bayes Assumption

- The Naïve Bayes classifier assumes that all features (or variables) in the data are conditionally independent given the class label
- Mathematically, for a set of features $X = x_1, x_2, \dots, x_n$ and a class $C$,

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

Now, by chain rule

$$P(X|C) = P(x_1, x_2, \dots, x_n|C\ ) = P(x_1|x_2, x_3..x_n, y).P(x_2|x_3, \dots x_n, y) \dots P(x_n|y)$$

but because of the Naive's conditional independence assumption

$$\boldsymbol{P(X|C) = P(x_1|C).P(x_2|C) \dots P(x_n|C)}$$

- Thus, by conditional independence, we have

$$\boldsymbol{P(C|X) = \frac{P(x_1|C).P(x_2|C) \dots P(x_n|C)\ .P(C)}{P(x_1).P(x_2) \dots P(x_n)}}$$

- And as denominator remains constant for all values, the posterior probability can then be:

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

# Assumption of Naive Bayes

- Feature independence
- Continuous features are normally distributed
- Discrete features have multinomial distributions
- Features are equally important
- No missing data

# Example 2

- Training Data Probabilities

| Apple (A)=0.5 | Orange (O)=0.5 |
|---|---|
| P(x1 =Red \| A)=0.8 | P(x 1 =Red \| O)=0.3 |
| P(x1 =Orange\|A)=0.2 | P(x 1 =Orange\|O)=0.7 |
| P(x 2 =Large \| A)=0.4 | P(x 2 =Large\|O)=0.6 |
| P(x 2 =Small \| A)=0.6 | P(x 2 =Small\|O)=0.4 |

- Question: A fruit is observed with the following characteristics: Color: Red ($x$1=Red), Size: Large ($x$ 2=Large). What is the classification of this fruit using Naïve Bayes?

# Solution 2

- Step 1: Compute Posterior for Apple $P(A|X)$

P(A|X)=P(A) × P(x1=red|A) × P(x2=large|A)=

$$0.5 \times 0.8 \times 0.4 = 0.16$$

- Step 2: Compute Posterior for Orange $P(\mathrm{O}|X)$

P(O|X)=P(O) × P(x1=red|O) × P(x2=large|O)=

$$0.5 \times 0.3 \times 0.6 = 0.09$$

P(A|X)>P(O|X), Apple

# Example 3

## 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 |

## Test Phase

– Given a new instance,

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

# Solution 3-Part1

- 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 |

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

# Solution 3-Part2

- Given a new instance,

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

- Look up tables

P(Outlook=*Sunny* | Play=*Yes*) = 2/9

P(Temperature=*Cool* | Play=*Yes*) = 3/9

P(Huminity=*High* | Play=*Yes*) = 3/9

P(Wind=*Strong* | Play=*Yes*) = 3/9

P(Play=*Yes*) = 9/14

P(Outlook=*Sunny* | Play=*No*) = 3/5

P(Temperature=*Cool* | Play==*No*) = 1/5

P(Huminity=*High* | Play=*No*) = 4/5

P(Wind=*Strong* | Play=*No*) = 3/5

P(Play=*No*) = 5/14

- MAP rule

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

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