

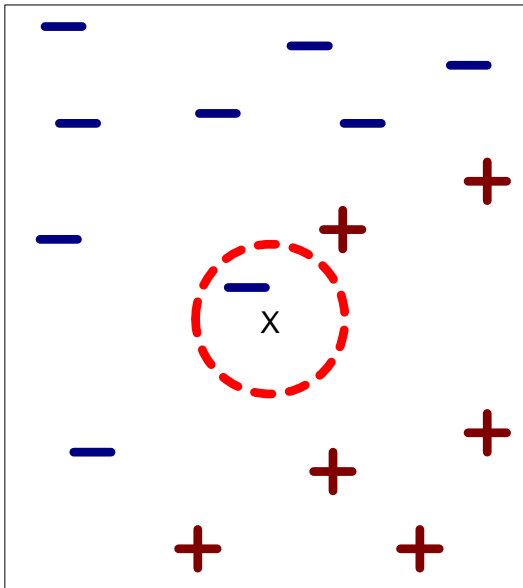
# Lecture outline

- Classification
- Naïve Bayes classifier
- Nearest-neighbor classifier

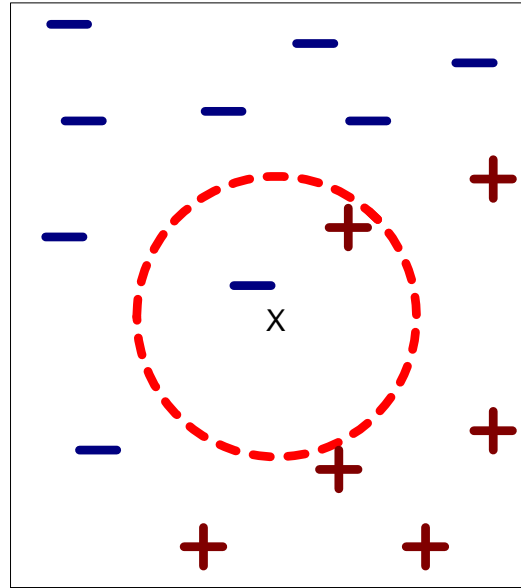
# Eager vs Lazy learners

- Eager learners: learn the model as soon as the training data becomes available
- Lazy learners: delay model-building until testing data needs to be classified
  - Rote classifier: memorizes the entire training data

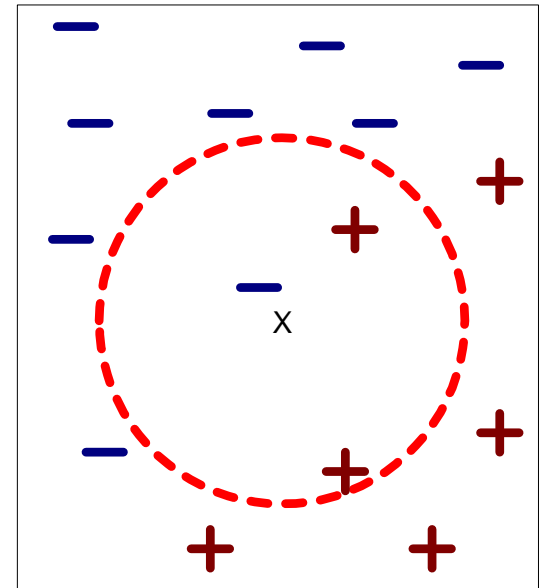
# k-nearest neighbor classifiers



(a) 1-nearest neighbor



(b) 2-nearest neighbor



(c) 3-nearest neighbor

**k**-nearest neighbors of a record  $x$  are data points that have the **k** smallest distance to  $x$

# **k**-nearest neighbor classification

- Given a data record **x** find its **k** closest points
  - Closeness: Euclidean, Hamming, Jaccard distance
- Determine the class of **x** based on the classes in the neighbor list
  - Majority vote
  - Weigh the vote according to distance
    - e.g., weight factor,  $w = 1/d^2$
  - Probabilistic voting

# Characteristics of nearest-neighbor classifiers

- Instance of *instance-based* learning
- No model building (lazy learners)
  - Lazy learners: computational time in classification
  - Eager learners: computational time in model building
- Decision trees try to find global models, k-NN take into account local information
- K-NN classifiers depend a lot on the choice of proximity measure

# Bayes Theorem

- **X, Y** random variables
- Joint probability:  **$\Pr(X=x, Y=y)$**
- Conditional probability:  **$\Pr(Y=y \mid X=x)$**
- Relationship between joint and conditional probability distributions

$$\Pr(X, Y) = \Pr(X \mid Y) \times \Pr(Y) = \Pr(Y \mid X) \times \Pr(X)$$

- **Bayes Theorem:**

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

# Bayes Theorem for Classification

- $X$ : attribute set
- $Y$ : class variable
- $Y$  depends on  $X$  in a *non-deterministic* way
- We can capture this dependence using

$\Pr(Y|X)$  : Posterior probability

**vs**

$\Pr(Y)$ : Prior probability

# Building the Classifier

- Training phase:
  - Learning the posterior probabilities  $\Pr(Y|X)$  for every combination of  $X$  and  $Y$  based on training data
- Test phase:
  - For test record  $X'$ , compute the class  $Y'$  that *maximizes the posterior probability*  
 $\Pr(Y'|X')$



# Bayes Classification: Example

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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

Figure 4.6. Training set for predicting borrowers who will default on loan payments.

$X' = (\text{Home Owner} = \text{No}, \text{Marital Status} = \text{Married}, \text{Annual Income} = 120\text{K})$

Compute:  $\Pr(\text{Yes} | X')$ ,  $\Pr(\text{No} | X')$  pick No or Yes with max Prob.

**How can we compute these probabilities??**

# Computing posterior probabilities

- Bayes Theorem

$$\Pr(Y | X) = \frac{\Pr(X | Y) \Pr(Y)}{\Pr(X)}$$

- **P(X)** is constant and can be ignored
- **P(Y)**: estimated from training data; compute the fraction of training records in each class
- **P(X|Y)?**

# Naïve Bayes Classifier

$$\Pr(X | Y = y) = \prod_{i=1}^d \Pr(X_i | Y = y)$$

- Attribute set  $\mathbf{X} = \{X_1, \dots, X_d\}$  consists of  $d$  attributes
- Conditional independence:
  - $\mathbf{X}$  conditionally independent of  $\mathbf{Y}$ , given  $\mathbf{Z}$ :  
 $\Pr(\mathbf{X} | \mathbf{Y}, \mathbf{Z}) = \Pr(\mathbf{X} | \mathbf{Z})$
  - $\Pr(\mathbf{X}, \mathbf{Y} | \mathbf{Z}) = \Pr(\mathbf{X} | \mathbf{Z}) \times \Pr(\mathbf{Y} | \mathbf{Z})$

# Naïve Bayes Classifier

$$\Pr(X | Y = y) = \prod_{i=1}^d \Pr(X_i | Y = y)$$

- Attribute set  $\mathbf{X} = \{X_1, \dots, X_d\}$  consists of  $d$  attributes

$$\Pr(Y | X) = \frac{\Pr(Y) \prod_{i=1}^d \Pr(X_i | Y)}{\Pr(X)}$$

# Conditional probabilities for categorical attributes

- Categorical attribute  $X_i$
- $\Pr(X_i = x_i | Y=y)$ : fraction of training instances in class  $y$  that take value  $x_i$  on the  $i$ -th attribute

$$\Pr(\text{homeOwner} = \text{yes} | \text{No}) = 3/7$$

$$\Pr(\text{MaritalStatus} = \text{Single} | \text{Yes}) = 2/3$$

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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

Figure 4.6. Training set for predicting borrowers who will default on loan payments.

# Estimating conditional probabilities for continuous attributes?

- Discretization?
- How can we discretize?

# Naïve Bayes Classifier: Example

- $X' = (\text{HomeOwner} = \text{No}, \text{MaritalStatus} = \text{Married}, \text{Income} = 120\text{K})$
- Need to compute  $\Pr(Y | X')$  or  $\Pr(Y) \times \Pr(X' | Y)$
- But  $\Pr(X' | Y)$  is
  - $Y = \text{No}$ :
    - $\Pr(\text{HO} = \text{No} | \text{No}) \times \Pr(\text{MS} = \text{Married} | \text{No}) \times \Pr(\text{Inc} = 120\text{K} | \text{No})$   
 $= 4/7 \times 4/7 \times 0.0072 = 0.0024$
  - $Y = \text{Yes}$ :
    - $\Pr(\text{HO} = \text{No} | \text{Yes}) \times \Pr(\text{MS} = \text{Married} | \text{Yes}) \times \Pr(\text{Inc} = 120\text{K} | \text{Yes})$   
 $= 1 \times 0 \times 1.2 \times 10^{-9} = 0$

# Naïve Bayes Classifier: Example

- $X' = (\text{HomeOwner} = \text{No}, \text{MaritalStatus} = \text{Married}, \text{Income} = 120\text{K})$
- Need to compute  $\Pr(Y | X')$  or  $\Pr(Y) \times \Pr(X' | Y)$
- But  $\Pr(X' | Y = \text{Yes})$  is **0**?
- Correction process:

$$\Pr(X_i = x_i | Y = y_j) = \frac{n_c + mp}{n + m}$$

$n_c$ : number of training examples from class  $y_j$  that take value  $x_i$

$n$ : total number of instances from class  $y_j$

$m$ : equivalent sample size (balance between prior and posterior)

$p$ : user-specified parameter (prior probability)



# Characteristics of Naïve Bayes Classifier

- Robust to isolated noise points
  - noise points are averaged out
- Handles missing values
  - Ignoring missing-value examples
- Robust to irrelevant attributes
  - If  $X_i$  is irrelevant,  $P(X_i|Y)$  becomes almost uniform
- Correlated attributes degrade the performance of NB classifier