Nearest Neighbor Classifier

- Given a test record, a simple approach to perform classification is to check whether the attributes of the record match those of one of the training examples exactly.
- The main problem with this approach is that some test records may not be classified since they do not match any training example.

Nearest Neighbor Classifier

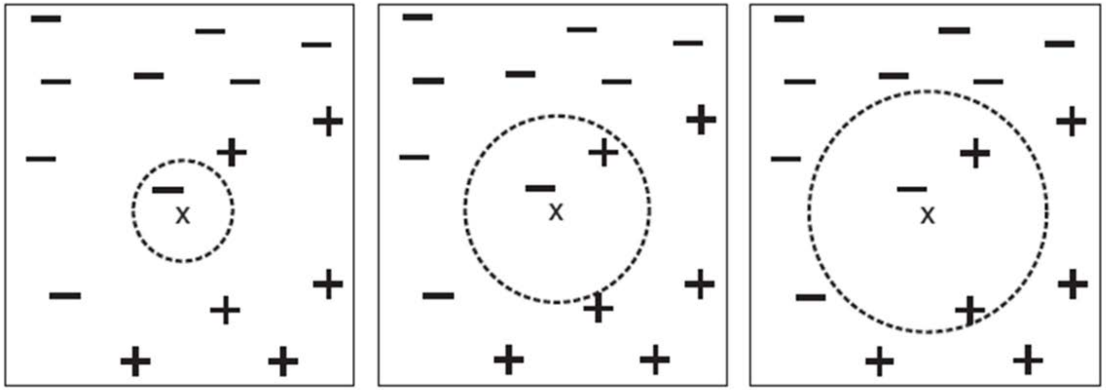
- To make this approach more flexible, we can find training examples with attributes that are relatively similar to those of the test example.
- These examples, known as nearest neighbors, can be used to determine the class label of the test example.

Nearest Neighbor Classifier

- A nearest neighbor classifier represents each example as a data point in an n-dimensional space, where n is the number of attributes.
- Given a test example, we calculate its distance to all the training examples using a suitable distance measure, and find its nearest neighbors.
- The k nearest neighbors of a given example refer to the k points that are closest to the example.

Nearest Neighbor Classifier

- The following figure illustrates the 1-, 2-, and 3-nearest neighbors of a data point located at the center of each circle.



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

- The data point is classified based on the class labels of its neighbors.

Nearest Neighbor Classifier

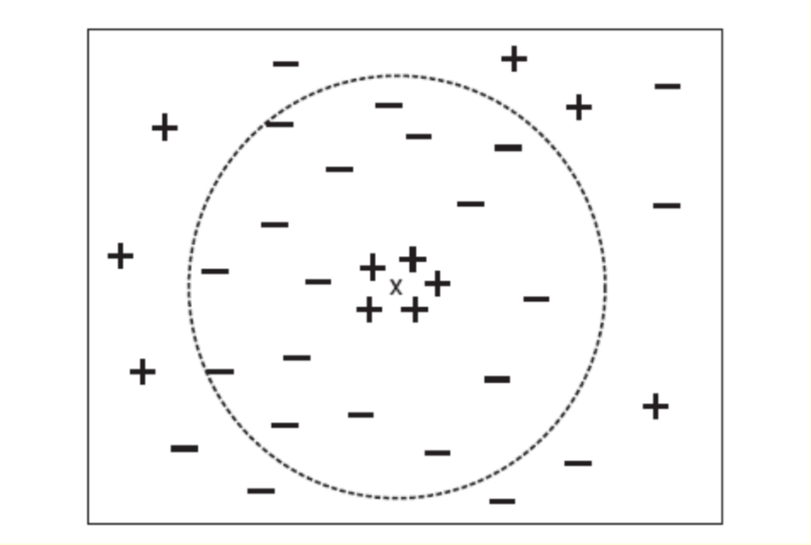
- When the nearest neighbors have more than one label, the data point is assigned to the majority class of its neighbors.
- In figure(a), the 1-nearest neighbor of the data point is a negative example.
- Therefore, the data point is assigned to the negative class.

Nearest Neighbor Classifier

- We now consider the example in figure(c), where the number of nearest neighbors is three.
- The neighborhood contains two positive examples and one negative example.
- Using the majority voting scheme, the data point is assigned to the positive class.
- When there is a tie between the classes, as shown in figure(b), we may randomly choose one of them to classify the data point.

Nearest Neighbor Classifier

- Choosing the right value for k is important.
- If k is too small, the presence of noise in the training data may affect the generalization performance.
- If k is too large, the classifier may misclassify the test instance, since the nearest neighbors may now include points that are far away from the instance.



Nearest Neighbor Classifier

- Initially, the algorithm computes the distance between each test example and all the training examples to determine its nearest neighbor list.
- Given this list, the test example is classified based on the majority class of its nearest neighbors.
- The computation of the nearest neighbor list can be costly if the number of training examples is large.
- Efficient indexing techniques could be applied to reduce the amount of computations for finding the nearest neighbors of a test example.

Probabilistic Classification Model

- In many applications the relationship between the attribute set and the class variable is non-deterministic.
- This could be due to noise or other confounding factors.
- A probabilistic classification model is required to estimate the probabilistic relationships between the attributes and the class variable.

Bayes Theorem

- Let X and Y be a pair of random variables.
- Their joint probability $P(X=x, Y=c)$ refers to the probability that
  - X will take on the value x, and
  - Y will take on the value c.
- The conditional probability $P(Y=c|X=x)$ refers to the probability that Y will take on the value c, given that X is observed to have the value x.

Bayes Theorem

- The joint and conditional probabilities for X and Y are related in the following way:

$$P(X,Y) = P(Y \mid X)P(X) = P(X \mid Y)P(Y)$$

- Rearranging the previous expression results in the Bayes theorem
  - $$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

Bayes Theorem for Classification

- Bayes theorem can be applied for classification by interpreting the attributes **X** and the class variable Y as random variables.
- Their probabilistic relationship can be captured using P(Y|**X**).
- This conditional probability is known as the posterior probability for Y.
- On the other hand, the probability P(Y) is known as the prior probability for Y.

Bayes Theorem for Classification

- During the training phase, we need to calculate a set of prior and class-conditional probabilities associated with different combinations of **X** and Y.
- Based on these probabilities, a test record **X** can be classified by finding the class Y that maximizes the posterior probability P(Y|**X**).

Bayes Theorem for Classification

- We consider the task of predicting whether a loan borrower will default on his/her payment.
- The training set with attributes Home Owner, Marital Status, and Annual Income is shown below:

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

Bayes Theorem for Classification

- Suppose we are given a test record with the following attribute set **X**
  - Home Owner = No
  - Marital Status = Married
  - Annual Income =$120K
- To classify the record, we need to compute the posterior probabilities P(Yes|**X**) and P(No|**X**) and compare their values.
  - If P(Yes|**X**)>P(No|**X**), the record is classified as Yes,
  - Otherwise, the record is classified as No.

Bayes Theorem for Classification

- The Bayes theorem for classification is expressed in the following form

$$P(Y \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid Y) P(Y)}{P(\mathbf{X})}$$

  where **X** denotes multiple attributes.
- The posterior probability P(Y|**X**) is expressed in terms of:
  - The prior probability P(Y),
  - The class-conditional probability P(**X**|Y), and
  - The evidence P(**X**)

Bayes Theorem for Classification

- When comparing the posterior probabilities for different Y values, P(**X**) is always constant and can be ignored.
- The prior probability can be estimated from the training set by computing the fraction of training records that belong to each class.
- To estimate the class-conditional probabilities P(**X**|Y), we consider the naïve Bayes classification approach.

Naïve Bayes Classifier

- A naïve Bayes classifier estimates the class-conditional probabilities by assuming that the attributes are conditionally independent, given the class label c.
- The conditional independence assumption can be expressed as follows:

$$P(\mathbf{X} \mid Y = c) = \prod_{u=1}^{n} P(X_u \mid Y = c)$$

where **X**=$\{X_1, X_2, \ldots\ldots, X_n\}$ consists of n attributes.

Naïve Bayes Classifier

- Based on this assumption, we need only to estimate the conditional probability of each X_u, given Y.
- This approach is more practical since it does not require a very large training set to obtain a good estimate of the probability.
- On the other hand, if we do not use this assumption, we need to compute the class-conditional probability for every combination of **X**.

Naïve Bayes Classifier

- To classify a test record **X**, the naïve Bayes classifier computes the posterior probability for each class Y:

$$P(Y \mid \mathbf{X}) = \frac{[\prod_{u=1}^{n} P(X_u \mid Y)]P(Y)}{P(\mathbf{X})}$$

- Since P(**X**) is fixed for every Y, it is sufficient to choose the class that maximizes the numerator term.

Categorical Attributes

- For a categorical attribute X_u, the conditional probability $P(X_u=x_u|Y=c)$ can be easily estimated.
- It is just the fraction of training instances in class c that takes on a particular attribute value x_u.

Categorical Attributes

- In the previous training set, three out of the seven people who repaid their loans also own a home.
- As a result, the conditional probability P(Home Owner=Yes|No) is equal to 3/7.
- Similarly, the conditional probability of defaulted borrowers who are single is given by P(Marital Status=Single|Yes)=2/3.

Continuous Attributes

- There are two ways to estimate the class-conditional probabilities for continuous attributes in naïve Bayes classifiers:
  - We can discretize each continuous attribute and then replace the continuous attribute value with its corresponding discrete interval.
  - We can assume a certain form of probability distribution for the continuous attribute.

Continuous Attributes

- A Gaussian distribution is usually chosen to represent the class-conditional probability for continuous attributes.
- For each class c_k, the class conditional probability for attribute X_u is:

$$P(X_u = x_u \mid Y = c_k) \approx \frac{\varepsilon}{\sqrt{2\pi}\sigma_{u,k}} \exp[-\frac{(x_u - \bar{x}_{u,k})^2}{2\sigma_{u,k}^2}]$$

- The distribution is characterized by two parameters:
  - Mean
  - Variance
- ε is a small constant.

Continuous Attributes

- $\bar{x}_{u,k}$ is the sample mean of X_u for all training records belonging to the class c_k.
- $\sigma^2_{u,k}$ is the sample variance of X_u of such training records.

Continuous Attributes

- For the Annual Income attribute, the parameters for the class No are estimated as follows:

$$\bar{x}_{\text{Annual Income, No}} = \frac{125+100+70+120+60+220+75}{7} = 110$$

$$\sigma^2_{\text{Annual Income, No}} = \frac{1}{6}[(125-110)^2+(100-110)^2+(70-110)^2+(120-110)^2+ (60-110)^2+(220-110)^2+(75-110)^2] = 2975$$

$$\sigma_{\text{Annual Income, No}} = \sqrt{2975} = 54.54$$

Continuous Attributes

- Given a test record with Annual Income equal to \$120K, the class-conditional probability can be computed as follows:

$$P(\text{Annual Income} = 120\text{K} \mid \text{No}) \approx \frac{\varepsilon}{\sqrt{2\pi}(54.54)} \exp[-\frac{(120-110)^2}{2(2975)}]$$

$$= 0.0072\varepsilon$$

Example

- Given the previous training set, we can estimate the followings:
  - The class-conditional probabilities for each categorical attribute
  - The sample mean and variance of the continuous attribute for each class.
- These values are summarized in the next slide.

Example

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

P(Home Owner=Yes|No) = 3/7
P(Home Owner=No|No) = 4/7
P(Home Owner=Yes|Yes) = 0
P(Home Owner=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No) = 1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/3
P(Marital Status=Divorced|Yes) = 1/3
P(Marital Status=Married|Yes) = 0

For Annual Income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

Example

- We consider the task of predicting the class label of a test record **X**
  - Home Owner = No
  - Marital Status = Married
  - Annual Income = $120K
- We need to compute the posterior probabilities P(No|**X**) and P(Yes|**X**).

Example

- The prior probability of each class can be estimated by calculating the fraction of training records that belong to each class.
- There are three records that belong to the class Yes, and seven records that belong to the class No.
- As a result, P(Yes)=3/10 and P(No)=7/10.

Example

- The class-conditional probabilities can be computed as follows:

$$
\begin{aligned}
P(\mathbf{X} \mid \text{No}) &= P(\text{Home Owner} = \text{No} \mid \text{No}) \times P(\text{Marital Status} = \text{Married} \mid \text{No}) \\
&\quad \times P(\text{Annual Income} = \$120\text{K} \mid \text{No}) \\
&\approx \frac{4}{7} \times \frac{4}{7} \times 0.0072\varepsilon = 0.0024\varepsilon
\end{aligned}
$$

$$
\begin{aligned}
P(\mathbf{X} \mid \text{Yes}) &= P(\text{Home Owner} = \text{No} \mid \text{Yes}) \times P(\text{Marital Status} = \text{Married} \mid \text{Yes}) \\
&\quad \times P(\text{Annual Income} = \$120\text{K} \mid \text{Yes}) \\
&\approx 1 \times 0 \times 1.2 \times 10^{-9}\varepsilon = 0
\end{aligned}
$$

Example

- The posterior probability for class No is

$$P(\text{No} \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \text{No})P(\text{No})}{P(\mathbf{X})} \approx \frac{0.0024\varepsilon \times \frac{7}{10}}{P(\mathbf{X})} = \frac{0.00168\varepsilon}{P(\mathbf{X})}$$

- The posterior probability for class Yes is

$$P(\text{Yes} \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \text{Yes})P(\text{Yes})}{P(\mathbf{X})} = \frac{0 \times \frac{3}{10}}{P(\mathbf{X})} = 0$$

- Since P(No|**X**)>P(Yes|**X**), the record is classified as No.