# DATA MINING 2
# Instance-based and Bayesian Classification

Riccardo Guidotti

a.a. 2019/2020

UNIVERSITÀ DI PISA

# Instance-based Classifiers

# Instance-based Classifiers

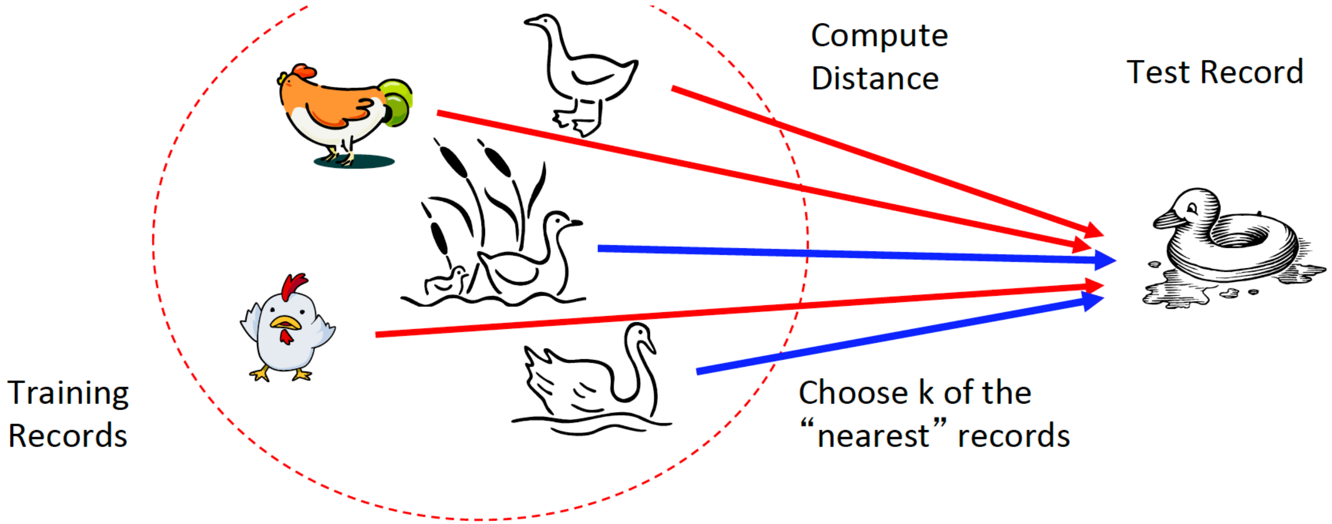
- Instead of performing explicit generalization, compare new instances with instances seen in training, which have been stored in memory.
- Sometimes called ***memory-based*** learning.
- ***Advantages***
  - Adapt its model to previously unseen data by storing a new instance or throwing an old instance away.
- ***Disadvantages***
  - Lazy learner: it does not build a model explicitly.
  - Classifying unknown records is relatively expensive: in the worst case, given $n$ training items, the complexity of classifying a single instance is $O(n)$.

# Nearest-Neighbor Classifier (K-NN)

Basic idea: If it walks like a duck, quacks like a duck, then it's probably a duck.

Requires three things

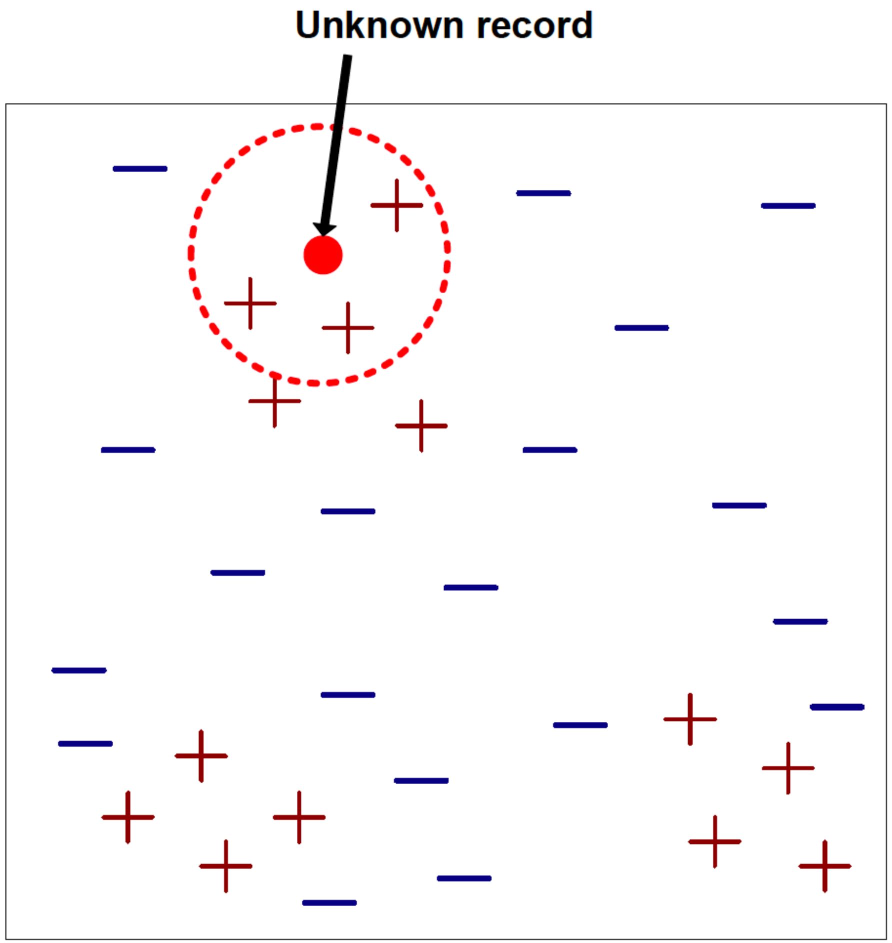
1. ***Training set*** of stored records
2. ***Distance metric*** to compute distance between records
3. ***The value of k***, the number of nearest neighbors to retrieve



# Nearest-Neighbor Classifier (K-NN)

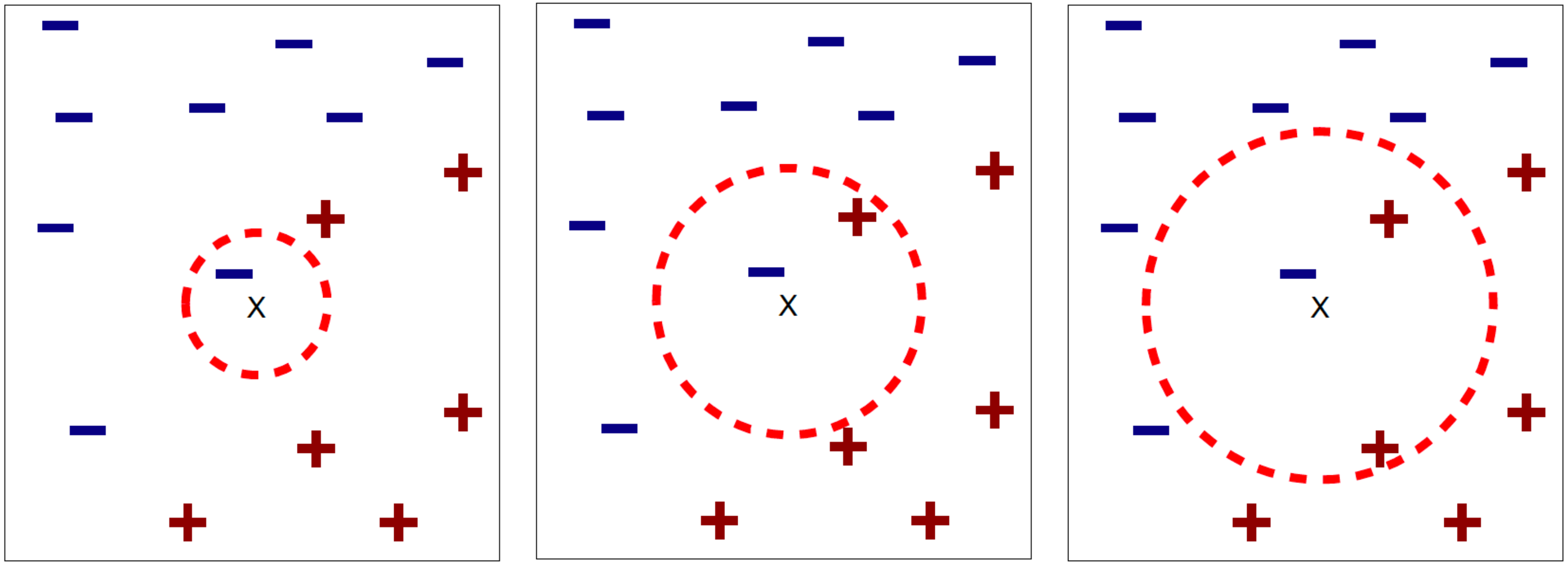
Given a set of training records (memory), and a test record:

1. ***Compute the distances*** from the records in the training to the test.
2. ***Identify the k "nearest" records***.
3. Use class labels of nearest neighbors to ***determine the class label*** of unknown record (e.g., by taking majority vote).



# Definition of Nearest Neighbor

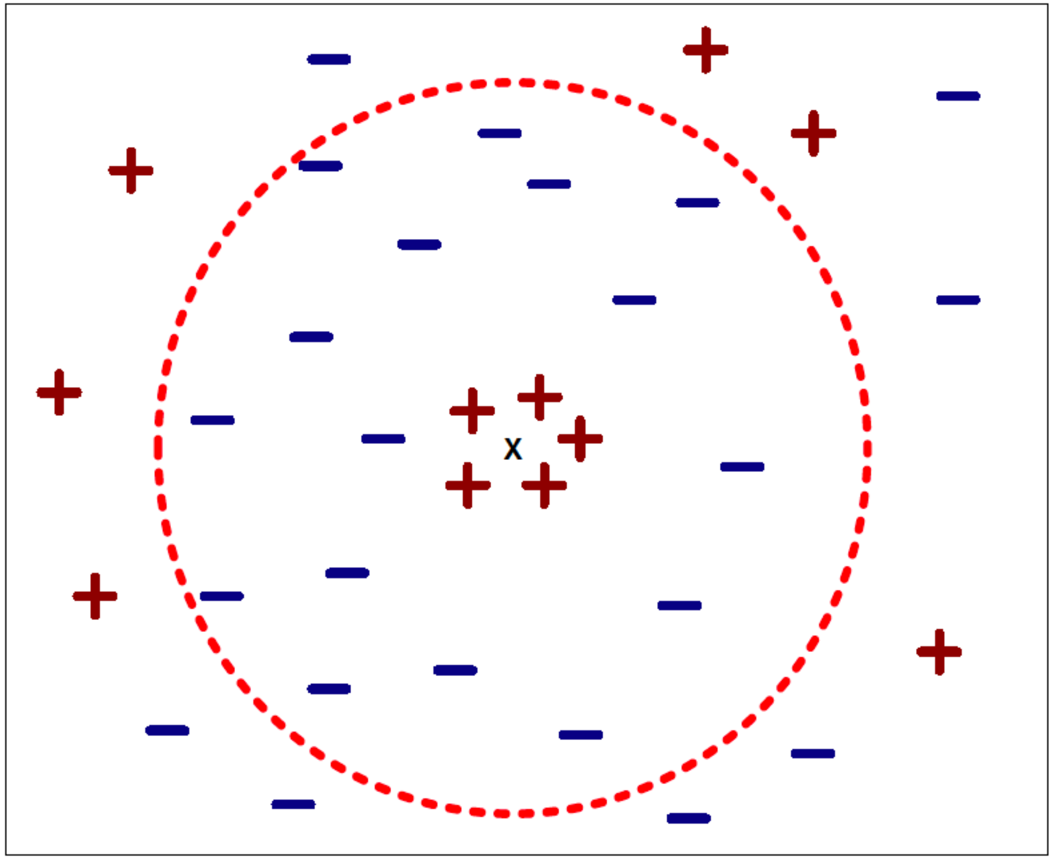
- *K*-nearest neighbors of a record *x* are data points that have the *k* smallest distance to *x*.




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

# Choosing the Value of K

- If k is too small, it is sensitive to noise points and it can leads to overfitting to the noise in the training set.
- If k is too large, the neighborhood may include points from other classes.
- General practice k = sqrt(N) where N is the number of samples in the training dataset.



# Nearest Neighbor Classification

Compute distance between two points:

- Euclidean distance $d(p, q) = \sqrt{\sum_i (p_i - q_i)^2}$

Determine the class from nearest neighbors

- take the majority vote of class labels among the k nearest neighbors
- weigh the vote according to distance (e.g. weight factor, w = $1/d^2$)

# Dimensionality and Scaling Issues

- Problem with Euclidean measure: high dimensional data can cause curse of dimensionality.
  - Solution: normalize the vectors to unit length

- Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes.
- Example:
  - height of a person may vary from 1.5m to 1.8m
  - weight of a person may vary from 10km to 200kg
  - income of a person may vary from $10K to $1M

# Parallel Exemplar-Based Learning System (PEBLS)

- PEBLS is a nearest-neighbor learning system (k=1) designed for applications where the instances have symbolic feature values.
- Works with both continuous and nominal features.
- For nominal features, the distance between two nominal values is computed using Modified Value Difference Metric (MVDM)
- $d(V_1, V_2) = \sum_i |\frac{n_{1i}}{n_1} - \frac{n_{2i}}{n_2}|$
- Where $n_1$ is the number of records that consists of nominal attribute value $V_1$ and $n_{1i}$ is the number of records whose target label is class $i$.

# Distance Between Nominal Attribute Values

- d(Status=Single, Status=Married) = | 2/4 – 0/4 | + | 2/4 – 4/4 | = 1
- d(Status=Single, Status=Divorced) = | 2/4 – 1/2 | + | 2/4 – 1/2 | = 0
- d(Status=Married, Status=Divorced) = | 0/4 – 1/2 | + | 4/4 – 1/2 | = 1
- d(Refund=Yes, Refund=No) = | 0/3 – 3/7 | + | 3/3 – 4/7 | = 6/7

| Class | Marital Status | | |
|---|---|---|---|
| | Single | Married | Divorced |
| Yes | 2 | 0 | 1 |
| No | 2 | 4 | 1 |

| Class | Refund | |
|---|---|---|
| | Yes | No |
| Yes | 0 | 3 |
| No | 3 | 4 |

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 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 |

# Distance Between Records

- $\delta(X, Y) = w_X w_Y \sum_{i=0}^{d} d(X_i, Y_i)$
- Each record *X* is assigned a weight $w_X = \frac{N_{X_{predict}}}{N_{X_{predict}^{correct}}}$, which represents its reliability
- $N_{X_{predict}}$ is the number of times *X* is used for prediction
- $N_{X_{predict}^{correct}}$ is the number of times the prediction using *X* is correct
- If $w_X \cong 1$ X makes accurate prediction most of the time
- If $w_X$> 1, then *X* is not reliable for making predictions. High $w_X$ >1 would result in high distance, which makes it less possible to use X to make predictions.

# Characteristics of Nearest Neighbor Classifiers

- Instance-based learner: makes predictions without maintaining abstraction, i.e., building a model like decision trees.
- It is a lazy learner: classifying a test example can be expensive because need to compute the proximity values between test and training examples.
- In contrast eager learners spend time in building the model but then the classification is fast.
- Make their prediction on local information and for low *k* they are susceptible to noise.
- Can produce wrong predictions if inappropriate distance functions and/or preprocessing steps are performed.

# Naïve Bayes Classifiers

# Bayes Classifier

- A probabilistic framework for solving classification problems.
- Let P be a probability function that assigns a number between 0 and 1 to events.
- $X = x$ an events is happening.
- $P(X = x)$ is the probability that events $X = x$.
- Joint Probability $P(X = x, Y = y)$
- Conditional Probability $P(Y = y \mid X = x)$
- Relationship: $P(X,Y) = P(Y|X)\ P(X) = P(X|Y)\ P(Y)$
- Bayes Theorem: $P(Y|X) = P(X|Y)P(Y) / P(X)$
- Another Useful Property: $P(X =x) = P(X=x, Y=0) + P(X=x, Y=1)$

# Bayes Theorem

- Consider a football game. Team 0 wins 65% of the time, Team 1 the remaining 35%. Among the game won by Team 1, 75% of them are won playing at home. Among the games won by Team 0, 30% of them are won at Team 1's field.
- If Team 1 is hosting the next match, which team will most likely win?
- Team 0 wins: $P(Y = 0) = 0.65$
- Team 1 wins: $P(Y = 1) = 0.35$
- Team 1 hosted the match won by Team 1: $P(X = 1|Y = 1) = 0.75$
- Team 1 hosted the match won by Team 0: $P(X = 1|Y = 0) = 0.30$
- Objective $P(Y = 1|X = 1)$

# Bayes Theorem

- $P(Y = 1|X = 1) = P(X = 1|Y = 1)P(Y = 1) / P(X = 1) =$
- $= 0.75 \times 0.35 / (P(X = 1, Y = 1) + P(X = 1, Y = 0))$
- $= 0.75 \times 0.35 / (P(X = 1|Y = 1)P(Y=1) + P(X = 1|Y = 0)P(Y=0))$
- $= 0.75 \times 0.35 / (0.75 \times 0.35 + 0.30 \times 0.65)$
- $= 0.5738$

- Therefore Team 1 has a better chance to win the match

# Bayes Theorem for Classification

- X denotes the attribute sets, $X = \{X_1, X_2, \dots X_d\}$
- Y denotes the class variable
- We treat the relationship probabilistically using $P(Y|X)$
- $$\underbrace{P(Y|X)}_{\text{Posterior Probability}} = \frac{\overbrace{P(X|Y)}^{\text{Likelihood}}\,\overbrace{P(Y)}^{\text{Prior Probability}}}{\underbrace{P(X)}_{\substack{\text{Evidence} \\ \text{(sum over alternative events)}}}}$$

# Bayes Theorem for Classification

- Learn the posterior P(Y | X) for every combination of X and Y.
- By knowing these probabilities, a test record X’ can be classified by finding the class Y’ that maximizes the posterior probability P(Y’|X’).
- This is equivalent of choosing the value of Y’ that maximizes P(X’|Y’)P(Y’).
- How to estimate it?

# Naïve Bayes Classifier

- It estimates the class-conditional probability by ***assuming that the attributes are conditionally independent*** given the class label y.
- The conditional independence is stated as:
- $P(X|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y)$
- where each attribute set X = $\{X_1, X_2, \ldots X_d\}$

# Conditional Independence

- Given three variables Y, $X_1$, $X_2$ we can say that Y is independent from $X_1$ given $X_2$ if the following condition holds:
- $P(Y \mid X_1, X_2) = P(Y|X_2)$

- With the conditional independence assumption, instead of computing the class-conditional probability for every combination of *X* we only have to estimate the conditional probability of each $X_i$ given *Y*.
- Thus, to classify a record the naive Bayes classifier computes the posterior for each class *Y* and takes the maximum class as result
- $P(Y|X) = P(Y) \prod_{i=1}^{d} P(X_i|Y = y) \,/ P(X)$

How to estimate ?

# How to Estimate Probability From Data

- Class $P(Y) = N_y / N$
- $N_y$ number of records with outcome y
- N number of records
- Categorical attributes
- $P(X = x \mid Y = y) = N_{xy} / N_y$
- $N_{xy}$ records with value x and outcome y

- P(Evade = Yes) = 3/10
- P(Marital Status = Single|Yes) = 2/3

| *Tid* | 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 |

# How to Estimate Probability From Data

Continuous attributes

- Discretize the range into bins
  - one ordinal attribute per bin
  - violates independence assumption
- Two-way split: $(X < v)$ or $(X > v)$
  - choose only one of the two splits as new attribute
- Probability density estimation:
  - Assume attribute follows a normal distribution
  - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
  - Once probability distribution is known, can use it to estimate the conditional probability $P(X|y)$

# How to Estimate Probability From Data

- Normal distribution
- $P(X_i = x_i \mid Y = y) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$
- $\mu_{ij}$ can be estimated as the mean of $X_i$ for the records that belongs to class $y_j$.
- Similarly, $\sigma_{ij}$ as the standard deviation.
- P(Income = 120|No) = 0.0072
  - mean = 110
  - std dev = 54.54

| *Tid* | 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 |

# Example

| Tid | 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 X = {Refund = No, Married, Income = 120k}

- P(Refund=Yes|No) = 3/7
- P(Refund=No|No) = 4/7
- P(Refund=Yes|Yes) = 0
- P(Refund=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) = 0/3

For taxable income:

- If class=No:
  - mean=110, variance=2975
- If class=Yes:
  - mean=90, variance=25

P(X|Class=No) = P(Refund=No|Class=No)
× P(Married| Class=No)
× P(Income=120K| Class=No)
= 4/7 × 4/7 × 0.0072
= 0.0024

P(X|Class=Yes) = P(Refund=No| Class=Yes)
× P(Married| Class=Yes)
× P(Income=120K| Class=Yes)
= 1 × 0 × 1.2 × 10-9
= 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)
=> Class = No

# M-estimate of Conditional Probability

- If one of the conditional probability is zero, then the entire expression becomes zero.
- For example, given X = {Refund = Yes, Divorced, Income = 120k}, if P(Divorced|No) is zero instead of 1/7, then
  - P(X|No) = 3/7 x 0 x 0.00072 = 0
  - P(X|Yes) = 0 x 1/3 x $10^{-9}$ = 0
- M-estimate P(X|Y) = $\frac{N_{xy}+mp}{N_y+m}$ (if P(X|Y) = $\frac{N_{xy}+1}{N_y+|Y|}$ is Laplacian estimation)
- m is a parameter, p is a user-specified parameter (e.g. probability of observing $x_i$ among records with class $y_j$.
- In the example with m = 3 and p = 1/m = 1/3 (i.e., Laplacian estimation) we have
- P(Married |Yes) = (0+3x1/3)/(3+3) = 1/6

# Naïve Bayes Classifier

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as Bayesian Belief Networks (BBN, not treated in this course)

# References

- Nearest Neighbor classifiers. Chapter 5.2. Introduction to Data Mining.
- Bayesian Classifiers. Chapter 5.3. Introduction to Data Mining.