כריית מידע
Data Mining (DM)

מרצה: ד"ר מרק לסט
מרכזת הקורס: ד"ר מיה הרמן

יחידה 7: למידה בייסיאנית ולמידה מבוססת תצפיות

תיאור היחידה

למידה בייסיאנית: משפט Bayes, אלגוריתם Naïve Bayes

למידה מבוססת תצפיות: שיטת k השכנים הקרובים ביותר (K-NN)

Lecture No. 7 – Bayesian Learning and Instance-based Learning

- Introduction to Bayesian Learning
- Naïve Bayes Algorithm
- Instance-based Learning: K-Nearest Neighbours Algorithm

Introduction to Bayesian Learning

- Basic Assumption
  - The observed data is governed by *probability distributions*
- Features
  - Using prior knowledge on probability distributions
  - Incremental learning of probabilities
    - Each observed example can incrementally increase or decrease the estimated probability
  - Probabilistic predictions of target values
  - Prediction by multiple hypotheses
  - A standard of **optimal decision making**
- Practical Algorithms
  - Naïve Bayes Classifier (comparable with decision tree classifiers)
  - Bayesian Belief Networks

Basic Formulas for Probabilities

- Product Rule
  - Probability $P(A \cap B)$ of a conjunction of two events A and B:
    - $P(A \cap B) = P(A / B) P(B) = P(B / A) P(A)$
- Sum Rule
  - Probability of a disjunction of two events A and B:
    - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Theorem of Total Probability
  - If events $A_1, \ldots, A_n$ are mutually exclusive with $\sum_i P(A_i) = 1$, then
    - $P(B) = \sum_i P(B/A_i)P(A_i)$

Bayesian Theorem: Basics

- Let **X** be a data sample ("*evidence*"): class label is unknown

age	income	student	credit_rating
<=30	high	no	fair
<=30	high	no	excellent
30…40	high	no	fair
>40	medium	no	fair
>40	medium	no	excellent

- Let H be a *hypothesis* that *X* belongs to class *C*
  - Optional classes: *Buys_Computer = Yes* and *Buys_Computer = No*
- Classification is to determine *P(H|**X**)*, the probability that the hypothesis holds given the observed data sample ***X***

Bayesian Theorem: Basics (cont.)

- *P(H)* (*prior probability*), the initial probability
  - E.g., **X** will buy computer, regardless of age, income, …
- *P(**X**):* probability that sample data is observed
  - E.g., the prob. that *X* is 31..40, medium income, etc.
- *P(**X**|H)* (*posteriori probability*), the probability of observing the sample ***X***, given that the hypothesis holds
  - E.g., Given *H* (***X*** will buy computer), the prob. that *X* is 31..40, medium income, etc.

Bayes Theorem

- Given training data *X*, posteriori probability of a hypothesis *H*, *P(H|X)* follows the Bayes theorem

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

- Example
  - *P(Buys_Computer = Yes / X is 31..40)=*

$$\frac{P(X \text{ is } 31..40 | Buys\_Computer = Yes) P(Buys\_Computer = Yes)}{P(X \text{ is } 31..40)}$$

- Informally, this can be written as
  - posterior =likelihood x prior / evidence

MAP (Maximum Posteriori) Hypothesis

- MAP (maximum posteriori) hypothesis

$$h_{MAP} \equiv \underset{h \in H}{\arg\max}\, P(h|D) = \underset{h \in H}{\arg\max}\, P(D|h)P(h).$$

- *D* – training data set
- Example
  - *X = 31..40*

$$h_{MAP} = \max\{P(31..40|Yes)P(Yes), P(31..40|No)P(No)\}.$$

- Practical difficulty: require initial knowledge of many probabilities, significant computational cost

Bayesian Theorem Example

Does patient have cancer or not?

Source: Mitchell, T.M., Machine Learning, McGraw-Hill, 1997

- ❖ A patient takes a lab test and the result comes back positive.
- ❖ It is known that the test returns a correct positive result in only 98% of the cases and a correct negative result in only 97% of the cases.
- ❖ Furthermore, only 0.008 of the entire population has this disease.

1. What is the probability that this patient has cancer?
2. What is the probability that he does not have cancer?
3. What is the diagnosis?

20595 כריית מידע
האוניברסיטה הפתוחה

Does patient have cancer or not? (cont'd)

- ❖ Medical diagnosis problem

  $P(cancer) = .008,$ $P(\neg cancer) = .992$

  $P(\oplus|cancer) = .98,$ $P(\ominus|cancer) = .02$

  $P(\oplus|\neg cancer) = .03$ $P(\ominus|\neg cancer) = .97$

- ❖ Maximum A Posteriori Hypothesis

  $P(\oplus|cancer)P(cancer) = (.98).008 = .0078$

  $P(\oplus|\neg cancer)P(\neg cancer) = (.03).992 = .0298$

  $h_{MAP} = \neg cancer$ Diagnosis

- ❖ Probability of Cancer

  $P(cancer|\oplus) = P(\oplus|cancer)P(cancer) / P(\oplus) =$

  $0.0078 / (0.0078 + 0.0298) = .21$

Bayes Optimal Classifier

- What is the most probable classification of the new instance given training data?
  - The most probable classification of the new instance is obtained by combining the predictions of all hypotheses, weighted by their posterior probabilities

$$\boxed{\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D)}$$

Bayes Optimal Classification Rule

  - *V* – set of possible classifications of the new instance ($v_j \in V$)
  - *D* – training data set
- No other classification method using the same hypothesis space and same prior knowledge can outperform this method <u>on average</u> (Mitchell, 1997)

Bayes Optimal Classifier(2)

- The *most probable classification* is different from *the most probable hypothesis*
- Example

MAP Hypothesis

$P(h_1|D) = .4$, $P(\Theta|h_1) = 0$, $P(\oplus|h_1) = 1$
$P(h_2|D) = .3$, $P(\Theta|h_2) = 1$, $P(\oplus|h_2) = 0$
$P(h_3|D) = .3$, $P(\Theta|h_3) = 1$, $P(\oplus|h_3) = 0$

Therefore $\sum_{h_i \in H} P(\Theta \mid h_i) P(h_i \mid D) = .6$

Most Probable Classification

$$\sum_{h_i \in H} P(\oplus \mid h_i) P(h_i \mid D) = .4$$

$$\arg\max_{v_j \in \{\oplus, \Theta\}} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D) = \Theta$$

Lecture No. 7 – Bayesian Learning and Instance-based Learning

- ❖ Introduction to Bayesian Learning
- ❖ Naïve Bayes Algorithm
- ❖ Instance-based Learning: K-Nearest Neighbours Algorithm

Towards Naïve Bayesian Classifier

- ❖ Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-D attribute vector **X** = $(x_1, x_2, \ldots, x_n)$
- ❖ Suppose there are *m* classes $C_1, C_2, \ldots, C_m$.
- ❖ Classification is to derive the maximum posteriori, i.e., the maximal $P(C_i|\mathbf{X})$
- ❖ This can be derived from Bayes' theorem

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- ❖ Since P(X) is constant for all classes, only

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

needs to be maximized

20595 כריית מידע
האוניברסיטה הפתוחה

Naïve Bayes Classifier (NBC)

- A simplified assumption: attributes are conditionally independent:
- The product of occurrence of say 2 elements x_1 and x_2, given the current class is C, is the product of the probabilities of each element taken separately, given the same class $P([y_1,y_2]/C) = P(y_1/C) * P(y_2/C)$
- No dependence relation between attributes
- Greatly reduces the computation cost, only count the class distribution.
- Once the probability $P(X|C_i)$ is known, assign X to the class with maximum $P(X|C_i)*P(C_i)$

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

$$C_{NB} = \arg\max_{C_i} \; P(C_i) * \prod_{k=1}^{n} P(x_k \mid C_i)$$

15 יחידה 7

20595 כריית מידע
האוניברסיטה הפתוחה

Training dataset – Example 1

Class:
C1:buys_computer= 'yes'
C2:buys_computer= 'no'

Data sample with unknown class:
X =(age<=30,
Income=medium,
Student=yes
Credit_rating= Fair)

age	income	student	credit_rating	s_comp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

16 יחידה 7

Naïve Bayesian Classifier: Example 1

- Compute $P(X/C_i)$ for each class

- P(age="<30" | buys_computer="yes") = 2/9=0.222
- P(age="<30" | buys_computer="no") = 3/5 =0.6
- P(income="medium" | buys_computer="yes")= 4/9 =0.444
- P(income="medium" | buys_computer="no") = 2/5 = 0.4
- P(student="yes" | buys_computer="yes)= 6/9 =0.667
- P(student="yes" | buys_computer="no")= 1/5=0.2
- P(credit_rating="fair" | buys_computer="yes")=6/9=0.667
- P(credit_rating="fair" | buys_computer="no")=2/5=0.4

- X=(age<=30 ,income =medium, student=yes, credit_rating=fair)

- $P(X|C_i)$:
  - P(X|buys_computer="yes")= 0.222 x 0.444 x 0.667 x 0.667 =0.044
  - P(X|buys_computer="no")= 0.6 x 0.4 x 0.2 x 0.4 =0.019
- $P(X|C_i)*P(C_i)$:
  - P(X|buys_computer="yes") * P(buys_computer="yes")=0.044 x 9/14 = 0.028
  - P(X|buys_computer="no") * P(buys_computer="no")=0.019 x 5/14 = 0.007
- **X belongs to class "buys_computer=yes"**

20595 כריית מידע
האוניברסיטה הפתוחה

Naïve Bayes Classifier (NBC) Example 2: Text Classification

Documents are classified as being scientific or commercial by the occurrence of the following three words: "paper", "research", and "product". The data obtained from 100 scientific documents and 100 commercial documents is summarized below:

Document	"Paper"	"Research"	"Product"
Scientific	80	90	20
Commercial	50	20	90

Explanation: 80 scientific documents included the word "paper", 90 commercial documents included the word "product", etc.

NBC Example 2: Text Classification Task

Classify the following text by using the Naïve Bayes algorithm:

The Oracle8i Appliance is a completely integrated database platform solution (based on the new Oracle8i Internet database) that combines all the necessary software components including the necessary operating environment. The **product** is entirely an Oracle-only solution which runs on Intel Architecture servers. In fact, since the Oracle8i Appliance only requires a few basic components from the operating system layer, there is no need for customers to access the operating system directly. All system functions and access will be provided via the Oracle Enterprise Manager management framework. It will be sold pre-configured and pre-installed by a wide variety of hardware vendors.

20595 כריית מידע
האוניברסיטה הפתוחה

NBC Example 2 (cont.)

Training Data

Document	"Paper"	"Research"	"Product"
Scientific	80	90	20
Commercial	50	20	90
Document	"Paper"	"Research"	"Product"
Scientific	0.8	0.9	0.2
Commercial	0.5	0.2	0.9

Test Data

	Apriori	"Paper"	"Research"	"Product"	Total	Rel.
P (Scientific)	0.5	0.2	0.1	0.2	**0.002**	**0.011**
P (Commercial)	0.5	0.5	0.8	0.9	**0.180**	**0.989**
					0.182	

Estimating Probabilities

- Two difficulties of estimating probability
  1. $\frac{n_c}{n}$ produces a biased underestimate of the probability
  2. When this probability estimate is zero, this probability term will dominate the Bayes classifier
  - Solution: using the m-estimate defined as follows
    m-estimate of probability: $$\frac{n_c + mp}{n + m}$$

    n_c : number of examples for which $v = v_j$ and $a = a_i$
    n : number of training examples for which $v = v_j$
    m : equivalent sample size
    Represents the reliability of the prior distribution
    p : uniform prior

  If m = 0, the m-estimate is equivalent to $\frac{n_c}{n}$

M-Estimate

- Ex. Suppose a dataset with 1000 tuples, income=low (0), income= medium (990), and income = high (10)
- Let equivalent sample size *m = 100*
- Uniform prior: 1/3
- *mp* = 33
- Use m- estimate

  Prob(income = low) = 33/1100
  Prob(income = medium) = (990+33)/1100
  Prob(income = high) = (10+33)/1100

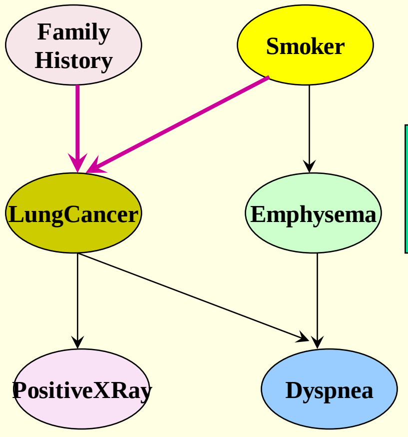
Laplacian Estimator

- Ex. Suppose a dataset with 1000 tuples, income=low (0), income= medium (990), and income = high (10)
- Use Laplacian correction (or Laplacian estimator)
  - Adding 1 to each case
    - Prob(income = low) = 1/1003
    - Prob(income = medium) = 991/1003
    - Prob(income = high) = 11/1003
  - The "corrected" prob. estimates are close to their "uncorrected" counterparts

Naïve Bayesian Classifier: Comments

- Advantages :
  - Easy to implement
  - Good results obtained in most of the cases
- Disadvantages
  - Assumption: class conditional independence , therefore loss of accuracy
  - Practically, dependencies exist among variables
  - E.g., hospitals: patients: Profile: age, family history etc
  - Symptoms: fever, cough etc., Disease: lung cancer, diabetes etc
  - Dependencies among these cannot be modeled by Naïve Bayesian Classifier
- How to deal with these dependencies?
  - Bayesian Belief Networks (beyond our scope)

Bayesian Belief Network: An Example



Bayesian Belief Networks

	(FH, S)	(FH, ~S)	(~FH, S)	(~FH, ~S)
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

The conditional probability table for the variable *LungCancer*: Shows the conditional probability for each possible combination of its parents

Lecture No. 7 – Bayesian Learning and Instance-based Learning

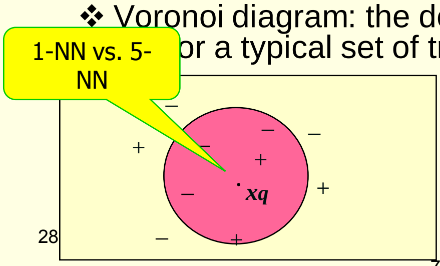
- Introduction to Bayesian Learning
- Naïve Bayes Algorithm
- Instance-based Learning: K-Nearest Neighbours Algorithm

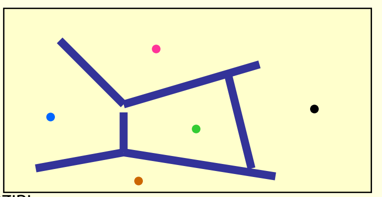
Instance-Based Methods

- Model-based ("eager") learning
  - Process training examples and store the model for classification of future instances
- Instance-based (memory-based) learning
  - Store training examples and delay the processing ("lazy evaluation") until a new instance must be classified
- Typical approaches of instance-based learning
  - *k*-nearest neighbor approach
    - Instances represented as points in a Euclidean space.
  - Kernel methods / Locally weighted regression
    - Construct local approximation
  - Case-based reasoning
    - Uses symbolic representations and knowledge-based inference

The k-Nearest Neighbor Algorithm

- All instances correspond to points in the n-D space.
- The nearest neighbors are defined in terms of Euclidean distance.
- The target function could be discrete- or real- valued.
- For discrete-valued, the k-NN returns the most common value among the *k* training examples nearest to x_q.
- For continuous-valued target functions, calculate the mean values of the *k* nearest neighbors
- Voronoi diagram: the decision surface induced by 1- or a typical set of training examples.





K-nearest neighbor for discrete classes

Algorithm (parameter *k*)

1. For each training example *(X,C(X))*
   add the example to our training list.

2. When a new example *Xq* arrives, assign class:

$C(Xq)$ = majority voting on the *k* nearest neighbors of *Xq*

$$C(Xq) = \text{argmax}\ v\ \Sigma_i\ \delta(v, C(Xi))$$

where $\delta(a,b) = 1$ if $a = b$ and 0 otherwise

K-nearest neighbor for real-valued functions

Algorithm (parameter *k*)

1. For each training example *(X,C(X))*
   add the example to our training list.

2. When a new example *Xq* arrives, assign class:

$C(Xq)$ = average value among *k* nearest neighbors of *Xq*

$$C(Xq) = \Sigma\ C(Xi)\ /\ k$$

20595 כריית מידע
האוניברסיטה הפתוחה

The Distance Between Examples

- We need a measure of distance in order to know who are the neighbours
- Assume that we have *T* attributes for the learning problem. Then one example point $\mathbf{x}$ has elements $x_t \in \Re$, $t=1,\ldots T$.
- The distance between two points $\mathbf{x}_i\,\mathbf{x}_j$ is often defined as the Euclidean distance:

$$d(\mathbf{x}_i,\mathbf{x}_j)=\sqrt{\sum_{t=1}^{T}[x_{ti}-x_{tj}]^2}$$

31 יחידה 7

20595 כריית מידע
האוניברסיטה הפתוחה

Euclidean Distance Illustration

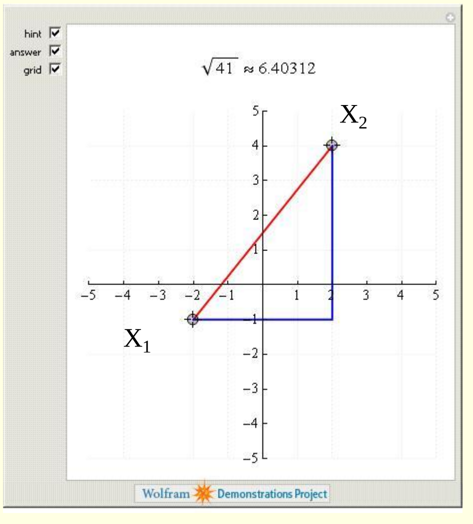
$X_1 = (-2, -1)$

$X_2 = (2, 4)$

Euclidean distance:

$(4^2+5^2)^{1/2} = 41^{1/2}$

$= 6.40312$



32 יחידה 7

20595 כריית מידע
האוניברסיטה הפתוחה

K-NN Example: Iris Dataset

- Full size: 150 observations
- Testing set: k = 50, 100, 150

Nearest Neighbor

k	sepal length in cm	sepal width in cm	petal length in cm	petal width in cm	Actual Class	**Dist (k,50)**	Predicted Class
8	5	3.4	1.5	0.2	1		
50	5	3.3	1.4	0.2		0.1414	1

Nearest Neighbor

k	sepal length in cm	sepal width in cm	petal length in cm	petal width in cm	Actual Class	**Dist (k,100)**	Predicted Class
97	5.7	2.9	4.2	1.3	2		
100	5.7	2.8	4.1	1.3		0.1414	2

Nearest Neighbor

k	sepal length in cm	sepal width in cm	petal length in cm	petal width in cm	Actual Class	**Dist (k,150)**	Predicted Class
128	6.1	3	4.9	1.8	3		
150	5.9	3	5.1	1.8		0.2828	3

33 יחידה 7

20595 כריית מידע
האוניברסיטה הפתוחה

When To Consider Nearest Neighbor

- Instances map to points in $\Re^n$
- Less than 20 attributes per instance
- Lots of training data
- Advantages
  - Training is very fast
  - Learn complex target functions
  - Don't lose information
- Disadvantages
  - Slow at query time
  - Limited interpretability
  - Curse of dimensionality
    - Distance between neighbors could be dominated by irrelevant attributes

34 יחידה 7

Summary

- Bayesian Learning provides a standard of optimal decision making
- Naïve Bayes Classifier is comparable with decision tree classifiers
- Instance-based (memory-based) methods delay the processing until a new instance must be classified
  - No classification model is produced