Classification Techniques: Bayesian Classification CSE450: Data Mining Summer 2018

SAH@DIU

Classification: 3 Step Process

- **1. Model construction (Learning):**
  - Each record (instance, example) is assumed to belong to a predefined class, as determined by one of the attributes
    - This attribute is called the target attribute
    - The values of the target attribute are the class labels
  - The set of all instances used for learning the model is called training set
- **2. Model Evaluation (Accuracy):**
  - Estimate accuracy rate of the model based on a test set
  - The known labels of test instances are compared with the predicts class from model
  - Test set is independent of training set otherwise over-fitting will occur
- **3. Model Use (Classification):**
  - The model is used to classify unseen instances (i.e., to predict the class labels for new unclassified instances)
  - Predict the value of an actual attribute

Classification Methods

- Decision Tree Induction
- Bayesian Classification
- K-Nearest Neighbor
- Neural Networks
- Support Vector Machines
- Association-Based Classification
- Genetic Algorithms
- Many More ….

- Also Ensemble Methods

Bayesian Learning

- **Bayes's theorem plays a critical role in probabilistic learning and classification**
  - Uses prior probability of each class given no information about an item
  - Classification produces a posterior probability distribution over the possible classes given a description of an item
  - The models are incremental in the sense that each training example can incrementally increase or decrease the probability that a hypothesis is correct. Prior knowledge can be combined with observed data
- **Given a data instance *X* with an unknown class label, *H* is the hypothesis that *X* belongs to a specific class *C***
  - The *conditional probability* of hypothesis *H* given *observation X,* Pr*(H|X),* follows the Bayes's theorem:

$$\boxed{\Pr(H \mid X) = \frac{\Pr(X \mid H)\Pr(H)}{\Pr(X)}}$$

- **Practical difficulty: requires initial knowledge of many probabilities**

Basic Concepts In Probability I

- **P(*A* | *B*) is the probability of *A* given *B***
- **Assumes that *B* is all and only information known.**

- **Note that:** $\boxed{P(A \wedge B) = P(A \mid B).P(B)}$

- **Bayes's Rule:**
  Direct corollary of above definition

$$\boxed{\begin{array}{l} P(A \wedge B) = P(A \mid B).P(B) \\ P(B \wedge A) = P(B \mid A).P(A) \\ \text{but } P(A \wedge B) = P(B \wedge A) \\ \therefore P(A \mid B).P(B) = P(B \mid A).P(A) \\ \therefore P(A \mid B) = \dfrac{P(B \mid A).P(A)}{P(B)} \end{array}}$$

- **Often written in terms of hypothesis and evidence:**

$$\boxed{P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}}$$

Basic Concepts In Probability II

- *A* and *B* are ***independent*** iff:

$$\boxed{P(A \mid B) = P(A)}$$

$$\boxed{P(B \mid A) = P(B)}$$

These two constraints are logically equivalent

- Therefore, if *A* and *B* are independent:

$$\boxed{P(A \mid B) = \frac{P(A \wedge B)}{P(B)} = P(A)} \qquad \boxed{P(A \wedge B) = P(A)P(B)}$$

Bayesian Classification

- Let set of classes be $\{c_1, c_2, \ldots c_n\}$
- Let E be description of an instance (e.g., vector representation)
- Determine class of E by computing for each class c_i

$$\boxed{P(c_i \mid E) = \frac{P(c_i)P(E \mid c_i)}{P(E)}}$$

- P(E) can be determined since classes are complete and disjoint:

$$\boxed{\sum_{i=1}^{n} P(c_i \mid E) = \sum_{i=1}^{n} \frac{P(c_i)P(E \mid c_i)}{P(E)} = 1}$$

$$\boxed{P(E) = \sum_{i=1}^{n} P(c_i)P(E \mid c_i)}$$

Bayesian Categorization (cont.)

- **Need to know:**
  - Priors: $P(c_i)$ and Conditionals: $P(E \mid c_i)$
- **$P(c_i)$ are easily estimated from data.**
  - If n_i of the examples in D are in c_i, then $P(c_i) = n_i / |D|$
- **Assume instance is a conjunction of binary features/attributes:**

$$\boxed{E = e_1 \wedge e_2 \wedge \cdots \wedge e_m}$$

Outlook	Tempreature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

$$\boxed{E = Outlook = rain \wedge Temp = cool \wedge Humidity = normal \wedge Windy = true}$$

Naïve Bayesian Classification

- **Problem: Too many possible combinations (exponential in *m*) to estimate all P(E | c_i)**

- **If we assume features/attributes of an instance are independent given the class (c_i) (*conditionally independent*)**

$$\boxed{P(E \mid c_i) = P(e_1 \wedge e_2 \wedge \cdots \wedge e_m \mid c_i) = \prod_{j=1}^{m} P(e_j \mid c_i)}$$

- **Therefore, we then only need to know P(e_j | c_i) for each feature and category**

Estimating Probabilities

- **Normally, probabilities are estimated based on observed frequencies in the training data.**
- **If D contains n_i examples in class c_i, and n_{ij} of these n_i examples contains feature/attribute e_j, then:**

$$\boxed{P(e_j \mid c_i) = \frac{n_{ij}}{n_i}}$$

- **If the feature is continuous-valued, P(e_j|c_i) is usually computed based on Gaussian distribution with a mean μ and standard deviation σ**

$$\boxed{g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

and P(e_j|c_i) is

$$\boxed{P(e_j \mid c_i) = g(e_j, \mu_{c_i}, \sigma_{c_i})}$$

Smoothing

- **Estimating probabilities from small training sets is error-prone:**
  - If due only to chance, a rare feature, e_k, is always false in the training data, $\forall c_i$:P(e_k | c_i) = 0.
  - If e_k then occurs in a test example, E, the result is that $\forall c_i$: P(E | c_i) = 0 and $\forall c_i$: P(c_i | E) = 0
- **To account for estimation from small samples, probability estimates are adjusted or smoothed**
- **Laplace smoothing using an m-estimate assumes that each feature is given a prior probability, *p*, that is assumed to have been previously observed in a "virtual" sample of size *m*.**

$$\boxed{P(e_j \mid c_i) = \frac{n_{ij} + mp}{n_i + m}}$$

- **For binary features, *p* is simply assumed to be 0.5.**

Naïve Bayesian Classifier - Example

Outlook	Tempreature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

- **Here, we have two classes C1="yes" (Positive) and C2="no" (Negative)**
- **Pr("yes") = instances with "yes" / all instances = 9/14**
- **If a new instance X had outlook="sunny", then Pr(outlook="sunny" | "yes") = 2/9 (since there are 9 instances with "yes" (or P) of which 2 have outlook="sunny")**
- **Similarly, for humidity="high", Pr(humidity="high" | "no") = 4/5**
- **And so on.**

Naïve Bayes (Example Continued)

- Now, given the training set, we can compute all the probabilities

Outlook	P	N
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5
Tempreature		
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	P	N
high	3/9	4/5
normal	6/9	1/5
Windy		
true	3/9	3/5
false	6/9	2/5

- Suppose we have new instance *X* = <sunny, mild, high, true>. How should it be classified?

$$X = <\text{sunny}, \text{mild}, \text{high}, \text{true}>$$

$$\Pr(X \mid \text{“no”}) = 3/5 \;.\; 2/5 \;.\; 4/5 \;.\; 3/5$$

- Similarly:

$$\Pr(X \mid \text{“yes”}) = (2/9 \;.\; 4/9 \;.\; 3/9 \;.\; 3/9)$$

Naïve Bayes (Example Continued)

- To find out to which class X belongs we need to maximize: **$Pr(X \mid C_i).Pr(C_i)$**, for each class C_i (here "yes" and "no")

$$X = \text{<sunny, mild, high, true>}$$

$$Pr(X \mid \text{"no"}).Pr(\text{"no"}) = (3/5 \,.\, 2/5 \,.\, 4/5 \,.\, 3/5) \,.\, 5/14 = 0.04$$

$$Pr(X \mid \text{"yes"}).Pr(\text{"yes"}) = (2/9 \,.\, 4/9 \,.\, 3/9 \,.\, 3/9) \,.\, 9/14 = 0.007$$

- To convert these to probabilities, we can normalize by dividing each by the sum of the two:
  - **Pr("no" | *X*) = 0.04 / (0.04 + 0.007) = 0.85**
  - **Pr("yes" | *X*) = 0.007 / (0.04 + 0.007) = 0.15**
- Therefore the new instance X will be classified as "no".

Text Naïve Bayes – Spam Example

Training Data

	t1	t2	t3	t4	t5	Spam
D1	1	1	0	1	0	no
D2	0	1	1	0	0	no
D3	1	0	1	0	1	yes
D4	1	1	1	1	0	yes
D5	0	1	0	1	0	yes
D6	0	0	0	1	1	no
D7	0	1	0	0	0	yes
D8	1	1	0	1	0	yes
D9	0	0	1	1	1	no
D10	1	0	1	0	1	yes

Term	P(t|no)	P(t|yes)
t1	1/4	4/6
t2	2/4	4/6
t3	2/4	3/6
t4	3/4	3/6
t5	2/4	2/6

P(no) = 0.4
P(yes) = 0.6

New email x containing t1, t4, t5 ➔ x = <1, 0, 0, 1, 1>

Should it be classified as spam = "yes" or spam = "no"?
Need to find P(yes | x) and P(no | x) …

Text Naïve Bayes - Example

Term	P(t|no)	P(t|yes)
t1	1/4	4/6
t2	2/4	4/6
t3	2/4	3/6
t4	3/4	3/6
t5	2/4	2/6

New email x containing t1, t4, t5

$$x = <1, 0, 0, 1, 1>$$

P(no) = 0.4
P(yes) = 0.6

$$P(yes \mid x) = [4/6 * (1-4/6) * (1-3/6) * 3/6 * 2/6] * P(yes) / P(x)$$
$$= [0.67 * 0.33 * 0.5 * 0.5 * 0.33] * 0.6 / P(x) = 0.11 / P(x)$$

$$P(no \mid x) = [1/4 * (1-2/4) * (1-2/4) * 3/4 * 2/4] * P(no) / P(x)$$
$$= [0.25 * 0.5 * 0.5 * 0.75 * 0.5] * 0.4 / P(x) = 0.019 / P(x)$$

To get actual probabilities need to normalize: note that P(yes | x) + P(no | x) must be 1

$0.11 / P(x) + 0.019 / P(x) = 1 \rightarrow P(x) = 0.11 + 0.019 = 0.129$

So:

$P(yes \mid x) = 0.11 / 0.129 = 0.853$

$P(no \mid x) = 0.019 / 0.129 = 0.147$