__/__/__

Ch-5 Classification - Alt. Techniques

- Rule-Based Classifier

→ It is a technique for classifying records using a collection of "if... then..." rules.

→ Rules for the model are represented in disjunctive normal form $R = (r_1 \vee r_2 \vee r_3 \ldots \vee r_K)$ where R = rule set and r_i are classification rules or disjunct.

$$r_i = \underbrace{(\text{Condition}_i)}_{\text{Rule Antecedent/ Precondition}} \rightarrow \underbrace{y_i}_{\text{Rule consequent (Predicted class)}}$$

$$\text{condition}_i = (A_1 \text{ op } v_1) \wedge (A_2 \text{ op } v_2) \wedge \ldots (A_k \text{ op } v_k)$$

where (A_j, v_j) is attribute value pair & op is logical operator $\{=, \neq, <, >, \leq, \geq\}$

→ A rule r covers a record x if the precondition of r matches the attributes of x. r is also said to be fired or triggered whenever it covers a given record.

→ Evaluating measures: Coverage and accuracy

→ The coverage of the rule r is defined as the fraction of records in data set D that trigger the rule r.

→ Accuracy or confidence factor is the fraction of records triggered by r whose class labels are equal to y.

$$\text{coverage}(r) = \frac{|A|}{|D|}$$

$$\text{Accuracy}(r) = \frac{|A \cap y|}{|A|}$$

where $|A|$ = no. of records that satisfy precondition

$|A \cap y|$ = no. " " " " antecedent & consequent

$|D|$ = total no. of records

→ Properties of Rule Based Classifier:

- Mutually Exhaustive Rules: No two rules in R are triggered by the same record. This ensures that every record is covered by almost one rule in R.
- Exhaustive Rules: There is a rule for each combination of attribute values. This ensures that every record is covered by atleast one rule in R.

→ If the rule set is not exhaustive, then a default rule, rd : ( ) → yd must be added to cover the remaining cases. It has an empty antecedent and is triggered when all other rules have failed. yd is default class and is typically assigned to majority class of training records not covered by existing rules.

→ If a rule set is not mutually exclusive then a record can be covered by more than one rule which can predict contradicting classes.

→ ways to overcome the problem of non-mutually exclusive rule sets:

① Ordered Rules: Rule sets are ordered in decreasing order of priority (eg. based on accuracy, coverage, total length desc. or the order in which they are generated). ordered Rule set known as Decision List. Record is classified by highest ranked rule. to avoid conflicting classes.

② Unordered Rules: The record triggered multiple rules and it is usually assigned the class that is predicted by most rules. It can also we weighted by rule's accuray.

- Advantages: Less susceptible to errors & less expensive than ordered set

→ Rule Ordering Schemes:

- Rule-Based ordering Scheme: orders the individual rules by some rule quality measure. Ensures every test records is classified by <u>best</u> rule covering it.

★ Drawback: Lower ranked rules are difficult to interpret because we have to assume the negation of prev. rules not ~~cover~~ triggered by record.

- Class Based Ordering Schemes: Rules that belong to same class appear together. Rules are then collectively sorted on the basis of their class information. Relative ordering among rules from same class is not important.

- Nearest Neighbor Classifier

→ Decision tree and rule based classifiers are eg. of <u>Eager learners</u> because they are designed to learn a model as soon as training data becomes available.

→ <u>Lazy learners</u> delay the process of ~~learning~~ modeling the training data until it is needed to classify the test egs.

→ <u>Rote Classifier</u> is a type of lazy learner, which memorize the entire training data & performs classification only if the test instance exactly matches.

★ Drawback: Records that do not match are not classified.

→ When we find all training examples relatively similar to test eg. then it is known as <u>nearest neighbor</u>

→ The data point is classified based on the class label of its neighbors. When there are more than one class label, majority class is assigned. When there is a tie, we can randomly choose one of them.

→ If k is too small: susceptible to overfitting bcoz of noise
If k is too large: may misclassify because it may include data points that are located too far.

__/__/__

→ Characteristics of Nearest-Neighbor Classifiers :

- It is part of instance-based learning technique which do not maintain model from data. They require proximity measure to determine similarity/distance & classification func. returns predicted class based on it.
- Do not req. model building. Expensive to compute proximity values individually b/w training & test eg. Eager learners spend most in model building
- Make predictions based on local info and hence susceptible for noise. Decision tree & Rule Based classifier find global model that fits entire input space
- Arbitrarily shaped decision boundaries, which provide more flexible model representation. Also has high variability ~~b~~because they depend on composition of training eg's. Increasing neighbors may reduce variability. Decision tree & rule based are often constrained to rectilinear decision boundaries
- Can produce wrong predictions unless appropriate proximity measure & data preprocessing steps are taken.

- Bayesian Classifiers

→ Class label of a test record cannot be predicted with certainity even though it's attribute set is identical to training eg's due to noisy data or certain confounding factors. So we ~~s~~ use probabilistic relationships.

→ Let X = attribute set & Y = class variable
if the class variable has non deterministic relationship with attributes then we can treat X & Y as random variables & capture their probabilistic r'ship using P(Y/X). This cond. prob. is also known as posterior prob. of Y as opp. to prior prob. P(Y)

→ During the traing phase, we need to learn the posterior prob. P(Y/X) for every comb. of X & Y based on training data which heps in classifying a test record x' by finding class Y' that maximizes the prob. $P(Y'/x')$

$$P(Y/X) = \frac{P(X/Y) \cdot P(Y)}{P(X)}$$

→ Naive Bayes Classifier

→ It estimates the class conditional probability by assuming that the attributes are conditionally independent, given the class label y.

Conditional independence assumption:

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

where $X = \{X_1, X_2, X_3 \ldots X_d\}$ d attributes

→ let X, Y & Z denote three sets of random variables. The variable X is said to be independent of Y, given Z if

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

→ The conditional independence of X & Y can also be written as

$$P(X, Y/Z) = \frac{P(X, Y, Z)}{P(Z)}$$

$$= \frac{P(X, Y, Z)}{P(\cancel{X}, \cancel{Z})} \times \frac{P(Y, Z)}{P(Z)}$$

$$= P(X \mid Y, Z) \times P(Y/Z)$$

$$= P(X/Z) \times P(Y/Z)$$

→ To classify a test record NBC computes the posterior prob.

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

since P(X) is fixed for every Y we have to maximize the numerator

__/__/____

- Alternative Metrics

→ For binary classification, the rare class is often denoted as the positive class, while the majority class is denoted as negative class.

		Predicted class	
		+	−
Actual	+	f_{++} (TP)	f_{+-} (FN)
class	−	f_{-+} (FP)	f_{--} (TN)

- True Positive (TP) or f_{++} are no. of + eg. correctly predicted as (+)
- False Negative (FN) or f_{+-} are " " (+) " wrongly " as (−)
- False Positive (FP) or f_{-+} " " " (−) eg. " " " (+)
- True Negative (TN) or f_{--} " " " (−) " correctly " " (−)

→ True Positive Rate (TPR) or sensitivity $= \frac{TP}{(TP+FN)}$

→ True Negative Rate (TNR) or specificity $= \frac{TN}{(TN+FP)}$

→ False Positive Rate (FPR) $= \frac{FP}{(FP+TN)}$

→ False Negative Rate (FNR) $= \frac{FN}{(FN+TP)}$

→ Precision, $p = \frac{TP}{FP+TP}$

→ Recall, $r = \frac{TP}{TP+FN}$

- Precision determines the fraction of records that actually turns out to be positive in the group the classifier declared as positive. The higher the precision, lower the no. of false (+) errors

__/__/__

- Recall measures the fraction of positive eg. correctly predicted as (+) by the classifier. Classes with large recall have very few (+) eg. misclassified as (-).

→ F_1 is the harmonic mean of precision & recall. Harmonic mean of two no's tends to be closer to the smaller of 2 no's & hence when F_1 is high we know that both p & r are high.

$$\boxed{F_1 = \frac{2}{\frac{1}{r}+\frac{1}{p}} = \frac{2rp}{r+p} = \frac{2\times TP}{2\times TP + FP + FN}}$$

→ F_β is the measure used to examine the tradeoff b/w r & p!

$$\boxed{F_\beta = \frac{(\beta^2+1)rp}{r+\beta^2 p} = \frac{(\beta^2+1)\times TP}{(\beta^2+1)TP + \beta^2 FP + FN}}$$

Both p & r are special cases of F_β by setting $\beta = 0$ & $\beta = \infty$ low values of β make F_β closer to precision & high value makes closer to recall.

Scanned with CamScanner