Classification Assessment

10/16

Let $\vec{x} = \{x_1, x_2, \dots, x_d\}^T \in \mathbb{R}^d$
$\hat{y} \in \{c_1, c_2, \dots, c_k\}$

Given a classifier $M: \mathbb{R}^d \to \{c_1, \dots, c_k\}$
created w/ training set

Assess M w/ **test set**, a set of pts w/ known labels
Let D be n pts in $\mathbb{R}^d$
w/ known labels
let y_i be the known label of $\vec{x}_i$
$\hat{y}_i = M(\vec{x}_i)$ is the predicted label

Performance Measures

Error rate

Let I be an indicator function
takes val 1 when arg is true false otherwise

$$\text{ErrorRate} = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

← - fraction of incorrect predictions
- est of prob of misclassification
- low error rate is better

Accuracy

$$\text{Accuracy} = 1 - \text{ErrorRate}$$

← fraction of correct predictions
est of the prob of correct classification
high accuracy is better

What could be a problem w/ just error rate or accuracy?

Contingency table based measures

Let D as our test set
$D_i \subset D$ w/ label c_i
$n_i = |D_i|$
$\mathcal{D} = \{D_1, \dots, D_k\}$

$R_j \subseteq D$ w/ predicted label c_i
$m_j = |R_j|$
$\mathcal{R} = \{R_1, \dots, R_k\}$

The **contingency table** N (also **Confusion matrix**)

is $k \times k$
w/ $N(i,j) = n_{ij} = |R_i \cap D_j| = |\{\vec{x}_l \in D \mid \hat{y}_l = c_i \text{ and } y_l = c_j\}|$

↑ # of pts w/ predicted label c_i and true label c_j

What is n_{ii}?

↳ # of pts w/ predicted label c_i and true label c_i ⇒ # of times we got c_i correct ✓

What is $\sum_{i=1, i \neq j}^{k} n_{ij}$?

↳ # of pts predicted label c_i actual label or c_j

# of times c_i wrong

Accuracy (Precision)

Class specific accuracy (precision)

$$acc_i = prec_i = \frac{n_{ii}}{m_i}$$

n_{ii} ← # of correct c_i's

m_i ← # w/ label c_i

higher is better

overall accuracy (precision)

$$\text{Accuracy} = \text{Precision} = \sum_{i=1}^{k} \frac{m_i}{n} acc_i = \frac{1}{n} \sum_{i=1}^{k} n_{ii}$$

Coverage / Recall

class specific coverage (recall)

$$\text{coverage}_i = \text{recall}_i = \frac{n_{ii}}{n_i}$$

higher is better

How can I get $\text{recall}_i = 1$ w/ little effort?

define $M(\vec{x}) = c_i \quad \forall \vec{x} \in \mathbb{R}^d$

Problem: precision would be very low

F-measure

class-specific F-measure : balance precision & recall

$$F_i = \frac{2}{\frac{1}{prec_i} + \frac{1}{rec_i}} = \frac{2n_{ii}}{n_i + m_i}$$ ← higher is better!