MISSING VALUES

- Ignore any record with a missing value?
- Easy for Naive Bayes
  - Don't include missing attribute in probability calculation
- Symbolic Attributes
  - Use 'Missing' as another value
  - Use the most common value [Weka]
  - Use 0 0 0 0 for 1 out of n coding [Neural Networks]
- Numeric Attributes
  - Use 0 [Could be dangerous]
  - Use 0 [Neural Networks, usually OK]
  - Use the average [Weka]
- Use machine learning techniques to predict missing value

COMPARISON CRITERIA FOR CLASSIFIERS

Accuracy Various measures

Speed Computational cost of learning and using

Robustness Deal with noisy data, missing values

Scalability Deal with large amounts of data

Interpretability Level of understanding and insight

	K-NN	J48
Accuracy	Similar	Similar
Speed Learning	Very Good	OK
Speed Using	Bad	Very Good
Robustness	OK	OK
Scalability	Very Bad	OK
Interpretability	Bad	Good

EVALUATION OF CLASSIFIERS

Building blocks of evaluation measures

Positive Tuples (P) The main class of interest, eg measles, sick, fraudulent

Negative Tuples (N) All others

True Positives (TP) Positive tuples correctly labelled by classifier

False Positives (FP) Negative tuples incorrectly labelled as positive

False Negatives (FN) Positive tuples incorrectly labelled as negative

P' Num of tuples labelled as positive, $(TP + FP)$

N' Num of tuples labelled as negative, $(TN + FN)$

Predicted Class

		yes	*no*	Total
Actual Class	*yes*	TP	FN	P
	no	FP	TN	N
	Total	P'	N'	P+N

MEASURES OF PERFORMANCE

Measure	Formula
Accuracy, Recognition Rate	$\frac{TP+TN}{P+N}$
Error Rate, Miscalssification Rate	$\frac{FP+FN}{P+N}$
Sensitivity, True Positive Rate, Recall	$\frac{TP}{P}$
Specificity, True Negative Rate	$\frac{TN}{N}$
Precision	$\frac{TN}{TP+FP}$
F, F_1, F-score Harmonic mean of precision and recall	$\frac{2\times precision\times recall}{precision+recall}$
AUC, Area under ROC curve	

Resubstitution Error Rate Training set error

Apparent Error Rate Training set error

SENSITIVITY AND SPECIFICITY

- Medical origin
- Confusion matrix for cancer classifier

Classes	Yes	No	Total	Recognition(%)
Yes	90	210	300	30.00
No	140	9560	9700	98.56
Total	230	9770	10000	96.40

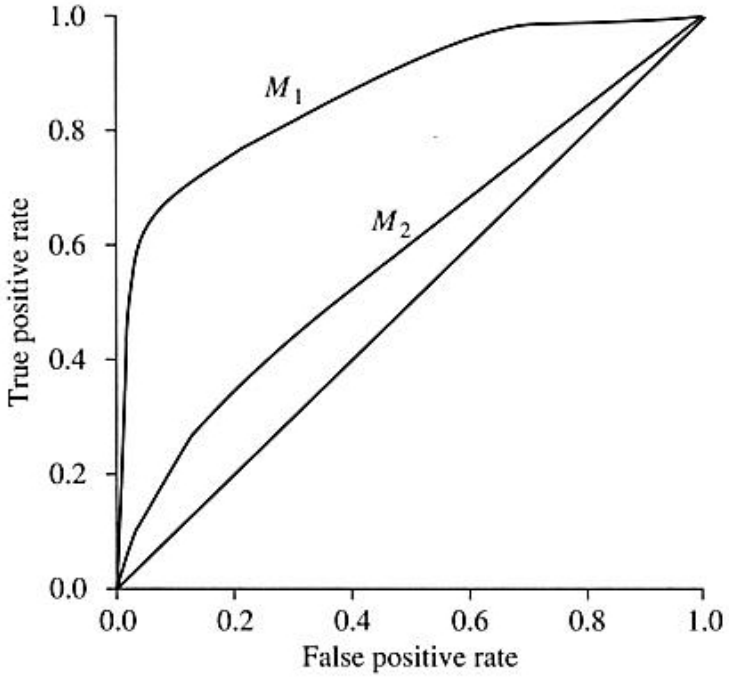
  Sensitivity: $90/300 = 30\%$

  Specificity: $9560/9700 = 98.56\%$

  Overall accuracy: $9650/10000 = 96.50\%$

- Classifier has high accuracy
- But low sensitivity, poor ability to label positive (rare) cases
- High specificity, good on negative cases
- Good enough?

ROC CURVE



- Receiver Operating Characteristic (ROC)
- From signal processing
- Recall Bayes classifier
  - Change the threshold, compute TP and FP
  - Plot TP vs FP
- Diagonal line is guessing
- M_1 is better than M_2 because for each FP value it gives more true positives
- What would be the perfect curve?
- Compare area under curve (AUC)

ROC CURVE

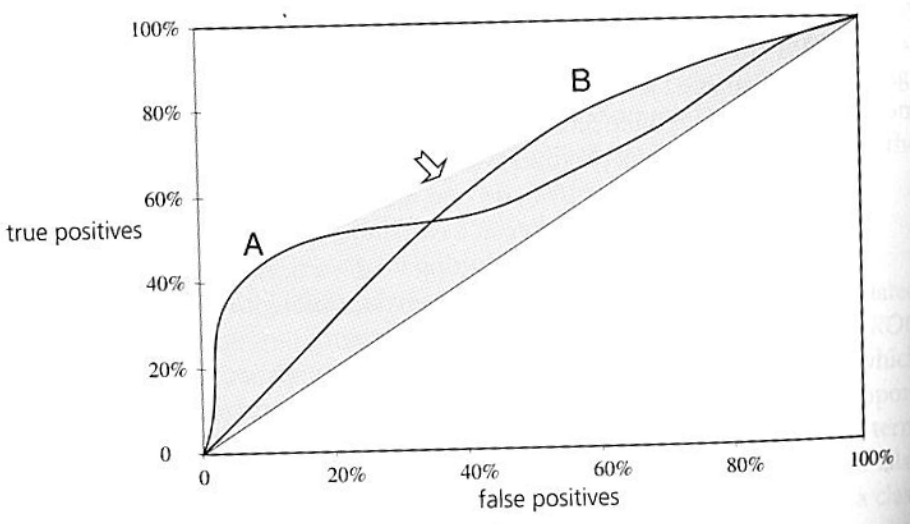


Figure 5.3 ROC curves for two learning methods.

- Which is better A or B?
- A is better if the aim is to cover around 40% of true positives
  - 5% false positive rate for A
  - 20% false positive rate for B
- B is better if the aim is to cover around 80% of true positives
  - 60% false positive rate for B
  - 80% false positive rate for A

NUMERIC PERFORMANCE MEASURES

Table 5.8 Performance measures for numeric prediction*.

Performance measure	Formula
mean-squared error	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$
root mean-squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}}$
mean absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{n}$
relative squared error	$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}$, where $\bar{a}=\frac{1}{n}\sum_i a_i$
root relative squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}}$
relative absolute error	$\frac{\lvert p_1-a_1\rvert+\ldots+\lvert p_n-a_n\rvert}{\lvert a_1-\bar{a}\rvert+\ldots+\lvert a_n-\bar{a}\rvert}$
correlation coefficient	$\frac{S_{PA}}{\sqrt{S_P S_A}}$, where $S_{PA}=\frac{\sum_i (p_i-\bar{p})(a_i-\bar{a})}{n-1}$, $S_p=\frac{\sum_i (p_i-\bar{p})^2}{n-1}$, and $S_A=\frac{\sum_i (a_i-\bar{a})^2}{n-1}$

* p are predicted values and a are actual values.

NUMERIC PERFORMANCE MEASURES

Mean Squared Error Most common, exaggerates outliers

Mean Absolute Error All errors treated evenly

Relative Squared Error Use if there is a large variation in values of target class

- Error of 50 in predicting 500
  - Error of .2 in predicting 2
  - are both 10% errors, but very different magnitude

Correlation coefficient Good performance means high correlation, near 1, between predicted and actual

NUMERIC PERFORMANCE MEASURES

Five performance measures on four classifiers of same dataset.

Which classifier is best?

Table 5.9 Performance measures for four numeric prediction models.

	A	B	C	D
root mean-squared error	67.8	91.7	63.3	57.4
mean absolute error	41.3	38.5	33.4	29.2
root relative squared error	42.2%	57.2%	39.4%	35.8%
relative absolute error	43.1%	40.1%	34.8%	30.4%
correlation coefficient	0.88	0.88	0.89	0.91

- Method D is best on all metrics
- Method C is second best
- Methods A and B inconclusive

Best method is generally the best no matter what measure is used

UNBALANCED DATA

- Class-imbalance problem: Rare positive example but numerous negative ones, e.g., medical diagnosis, fraud, oil-spill, fault, etc.
- Traditional methods assume a balanced distribution of classes and equal error costs: not suitable for class-imbalanced data
- Typical methods for unbalanced data in 2-class classification
  - Cost based learning and classification
  - Oversampling: re-sampling of data from positive class
  - Under-sampling: randomly eliminate examples from negative class
  - Threshold-moving: moves the decision threshold, t, to pick up the rare class examples
  - Use Ensemble techniques
  - Multi-objective Evolutionary classifier, sensitivity vs specificity
- No good solutions for multiclass problems
- Still a major research issue