Performance Measures for Machine Learning

1

Performance Measures

- Accuracy
- Weighted (Cost-Sensitive) Accuracy
- Lift
- ROC
  - ROC Area
- Precision/Recall
  - F
  - Break Even Point
- Similarity of Various Performance Metrics via MDS (Multi-Dimensional Scaling)

2

Accuracy

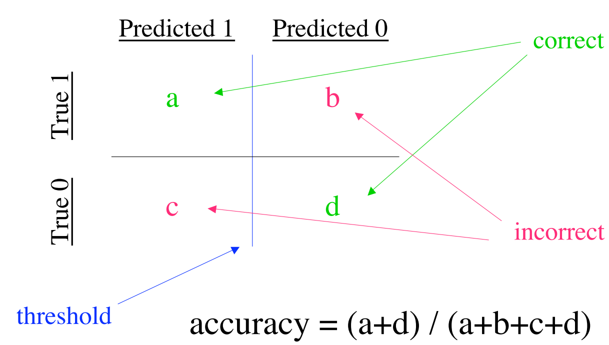
- Target: 0/1, -1/+1, True/False, …
- Prediction = f(inputs) = f(x): 0/1 or Real
- Threshold: f(x) > thresh => 1, else => 0
- If threshold(f(x)) and targets both 0/1:

$$accuracy = \frac{\sum_{i=1\ldots N}\left(1 - \left|target_i - threshold(f(\vec{x}_i))\right|_{ABS}\right)}{N}$$

- #right / #total
- p("*correct*"): p(threshold(f(x)) = target)

3

Confusion Matrix



accuracy = (a+d) / (a+b+c+d)

4

	Predicted 1	Predicted 0
True 1	true positive	false negative
True 0	false positive	true negative

	Predicted 1	Predicted 0
True 1	TP	FN
True 0	FP	TN

	Predicted 1	Predicted 0
True 1	hits	misses
True 0	false alarms	correct rejections

	Predicted 1	Predicted 0
True 1	$P(pr1 \mid tr1)$	$P(pr0 \mid tr1)$
True 0	$P(pr1 \mid tr0)$	$P(pr0 \mid tr0)$

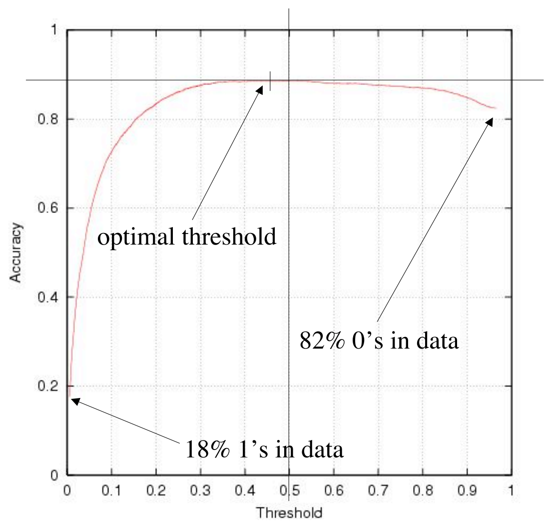
Prediction Threshold

	Predicted 1	Predicted 0
True 1	0	b
True 0	0	d

- threshold > MAX(f(x))
- all cases predicted 0
- (b+d) = total
- accuracy = %False = %0's

	Predicted 1	Predicted 0
True 1	a	0
True 0	c	0

- threshold < MIN(f(x))
- all cases predicted 1
- (a+c) = total
- accuracy = %True = %1's



Problems with Accuracy

- Assumes equal cost for both kinds of errors
  - cost(b-type-error) = cost (c-type-error)
- is 99% accuracy good?
  - can be excellent, good, mediocre, poor, terrible
  - depends on problem
- is 10% accuracy bad?
  - information retrieval
- BaseRate = accuracy of predicting predominant class
  (on most problems obtaining BaseRate accuracy is easy)

Percent Reduction in Error

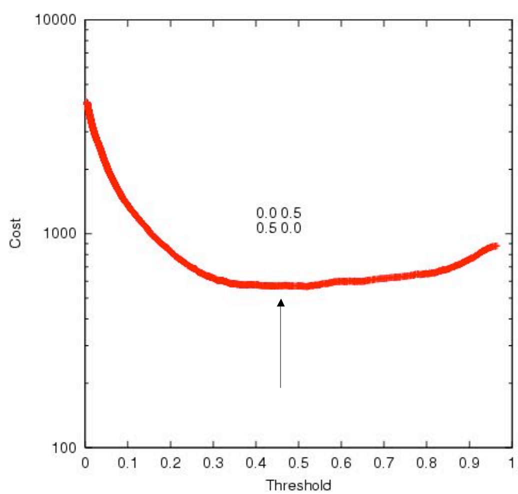
- 80% accuracy = 20% error
- suppose learning increases accuracy from 80% to 90%
- error reduced from 20% to 10%
- 50% reduction in error

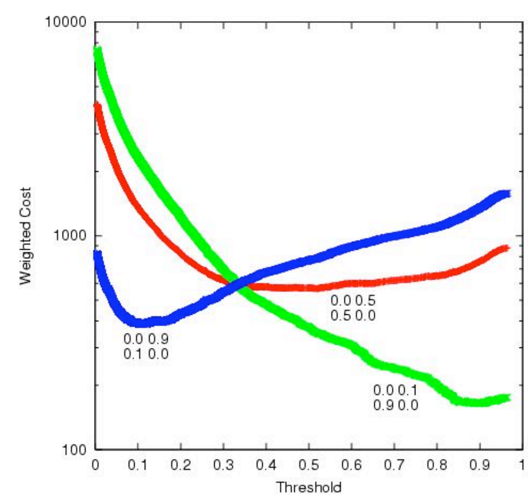
- 99.90% to 99.99% = 90% reduction in error
- 50% to 75% = 50% reduction in error
- can be applied to many other measures

Costs (Error Weights)

	Predicted 1	Predicted 0
True 1	w_a	w_b
True 0	w_c	w_d

- Often $W_a = W_d =$ **zero** and $W_b \neq W_c \neq$ **zero**





Lift

- not interested in accuracy on entire dataset
- want accurate predictions for 5%, 10%, or 20% of dataset
- don't care about remaining 95%, 90%, 80%, resp.
- typical application: marketing

$$lift(threshold) = \frac{\%\,positives > threshold}{\%\,dataset > threshold}$$

- how much better than random prediction on the fraction of the dataset predicted true (f(x) > threshold)

13

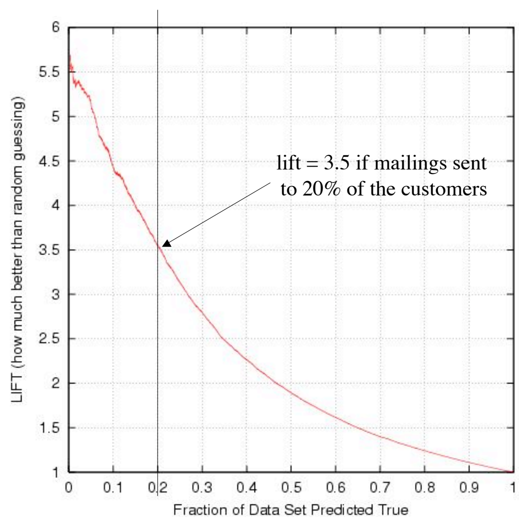
Lift

	Predicted 1	Predicted 0
True 1	a	b
True 0	c	d

threshold

$$lift = \frac{a/(a+b)}{(a+c)/(a+b+c+d)}$$

14

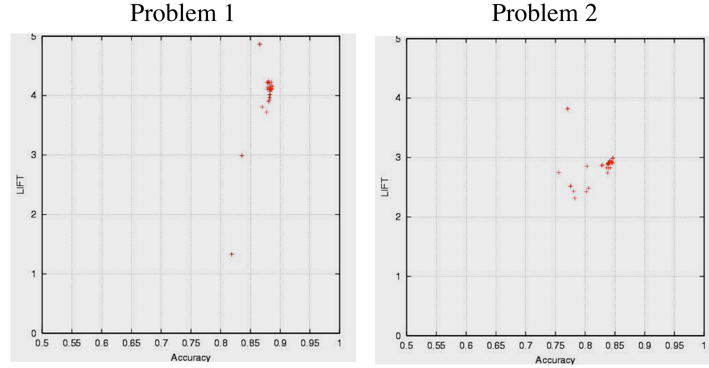


15

Lift and Accuracy do not always correlate well

Problem 1

Problem 2



(thresholds arbitrarily set at 0.5 for both lift and accuracy)

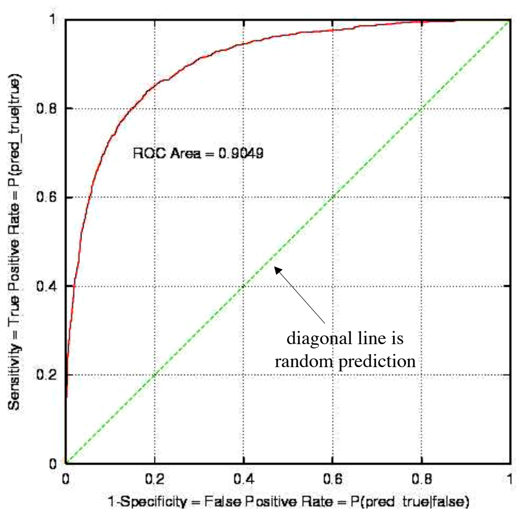
16

ROC Plot and ROC Area

- Receiver Operator Characteristic
- Developed in WWII to statistically model false positive and false negative detections of radar operators
- Better statistical foundations than most other measures
- Standard measure in medicine and biology
- Becoming more popular in ML

ROC Plot

- Sweep threshold and plot
  - TPR vs. FPR
  - Sensitivity vs. 1-Specificity
  - P(true|true) vs. P(true|false)
- Sensitivity = a/(a+b) = LIFT numerator = Recall (see later)
- 1 - Specificity = 1 - d/(c+d)



Properties of ROC

- ROC Area:
  - 1.0: perfect prediction
  - 0.9: excellent prediction
  - 0.8: good prediction
  - 0.7: mediocre prediction
  - 0.6: poor prediction
  - 0.5: random prediction
  - <0.5: something wrong!

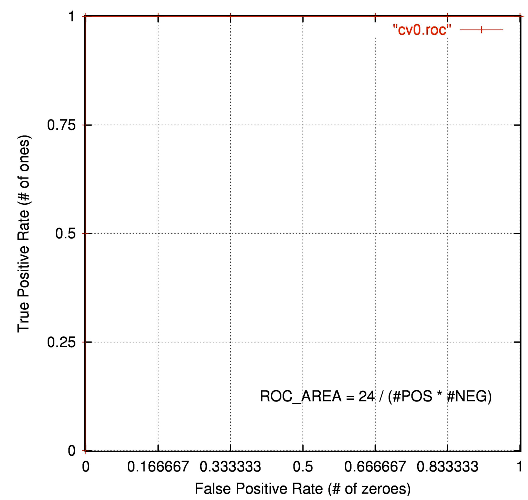
Wilcoxon-Mann-Whitney

$$ROCA = 1 - \frac{\#\_pairwise\_inversions}{\#POS * \#NEG}$$

where

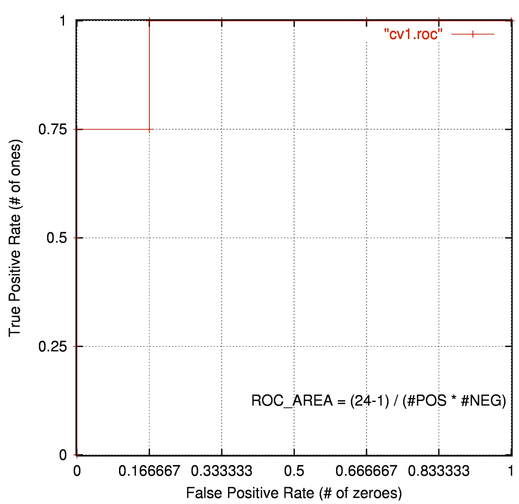
$$\#\_pair\_inversions = \sum_{i,j} I\left[\left(P(x_i) > P(x_j)\right) \,\&\, \left(T(x_i) < T(x_j)\right)\right]$$

21



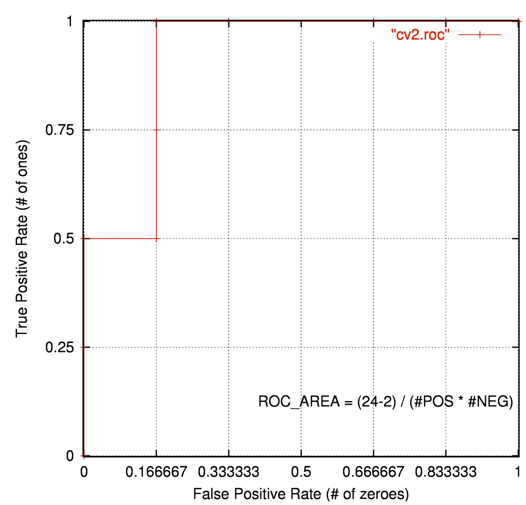
1	0.9
1	0.8
1	0.7
1	0.6
0	0.5
0	0.4
0	0.3
0	0.3
0	0.2
0	0.1

22



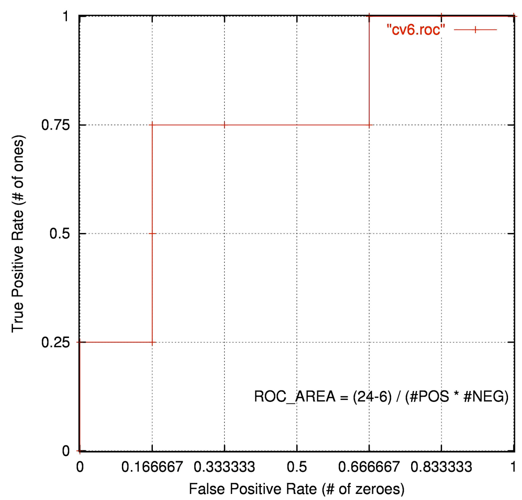
1	0.9
1	0.8
1	0.7
1	0.6
0	0.65
0	0.4
0	0.3
0	0.3
0	0.2
0	0.1

23



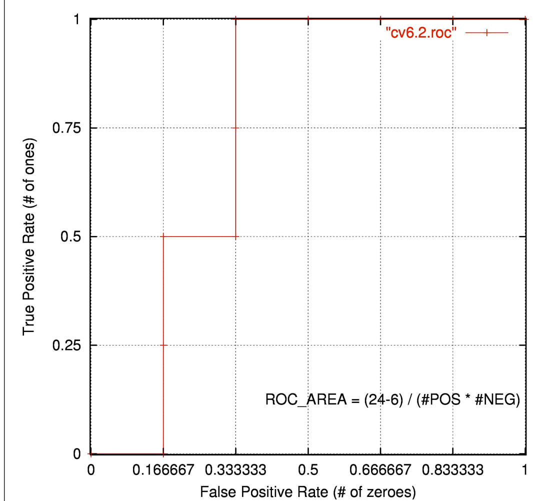
1	0.9
1	0.8
1	0.7
1	0.6
0	0.75
0	0.4
0	0.3
0	0.3
0	0.2
0	0.1

24



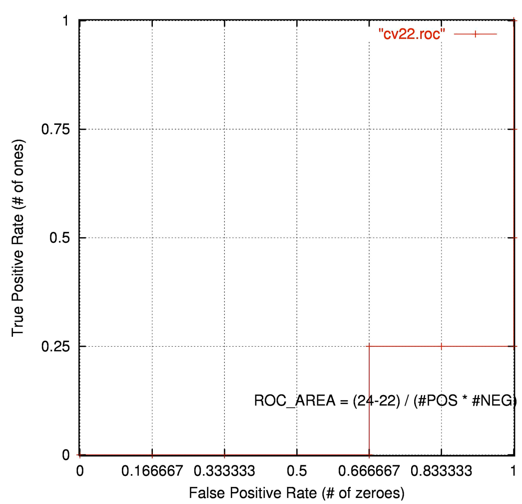
1	0.9
1	0.25
1	0.7
1	0.6
0	0.75
0	0.4
0	0.3
0	0.3
0	0.2
0	0.1

25



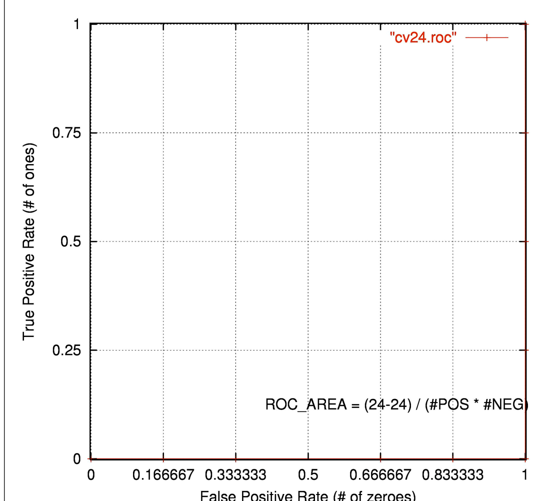
1	0.9
1	0.8
1	0.7
1	0.6
0	0.75
0	0.4
0	0.3
0	0.93
0	0.2
0	0.1

26



1	0.09
1	0.25
1	0.07
1	0.06
0	0.75
0	0.4
0	0.3
0	0.3
0	0.2
0	0.1

27



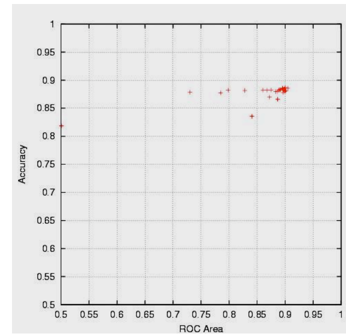
1	0.09
1	0.025
1	0.07
1	0.06
0	0.75
0	0.4
0	0.3
0	0.3
0	0.2
0	0.1

28

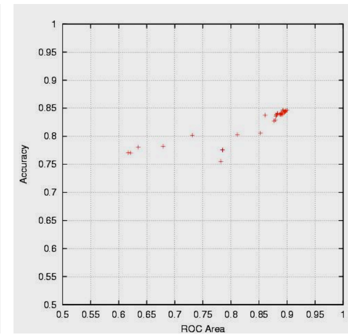
Properties of ROC

- Slope is non-increasing
- Each point on ROC represents different tradeoff (cost ratio) between false positives and false negatives
- Slope of line tangent to curve defines the cost ratio
- ROC Area represents performance averaged over all possible cost ratios
- If two ROC curves do not intersect, one method dominates the other
- If two ROC curves intersect, one method is better for some cost ratios, and other method is better for other cost ratios

Problem 1



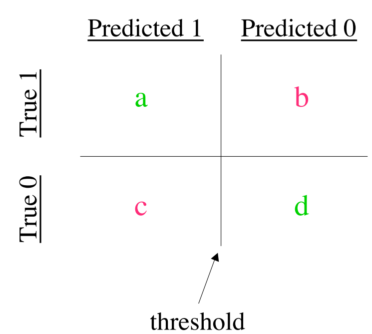
Problem 2



Precision and Recall

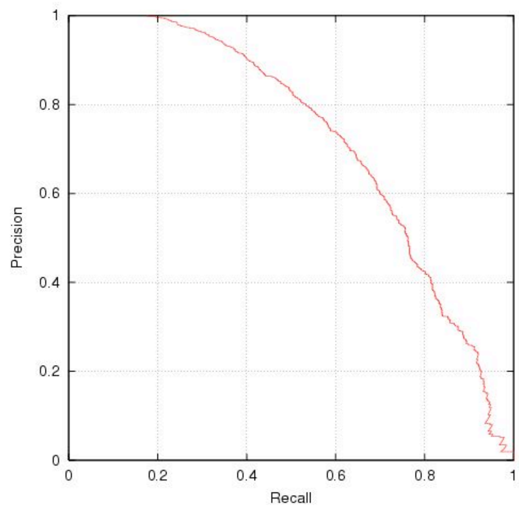
- typically used in document retrieval
- Precision:
  - how many of the returned documents are correct
  - precision(threshold)
- Recall:
  - how many of the positives does the model return
  - recall(threshold)
- Precision/Recall Curve: sweep thresholds

Precision/Recall



$$PRECISION = a/(a+c)$$

$$RECALL = a/(a+b)$$



Summary Stats: F & BreakEvenPt

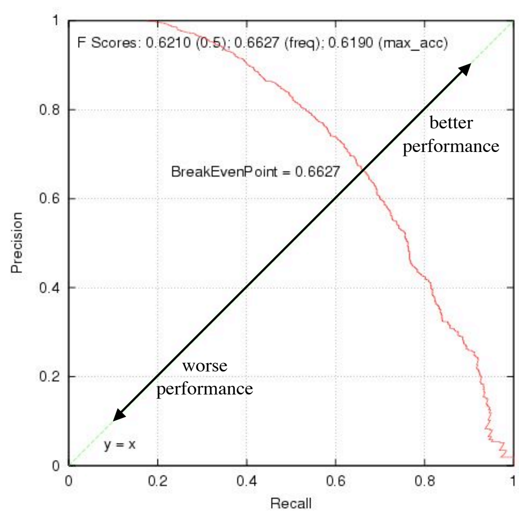
$$PRECISION = a/(a+c)$$

$$RECALL = a/(a+b)$$

$$F = \frac{2 * (PRECISION \times RECALL)}{(PRECISION + RECALL)}$$

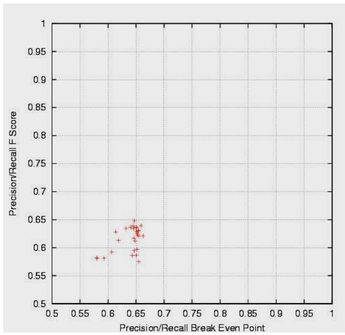
harmonic average of precision and recall

$$BreakEvenPoint = PRECISION = RECALL$$

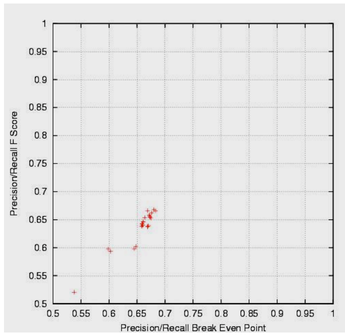


F and BreakEvenPoint do not always correlate well

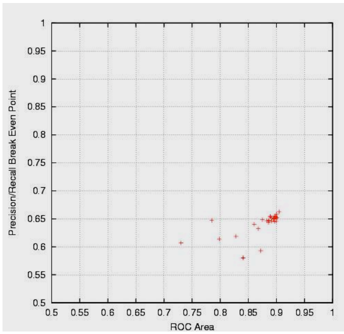
Problem 1



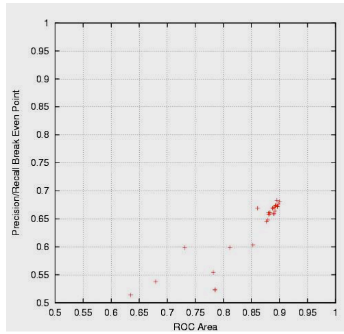
Problem 2



Problem 1

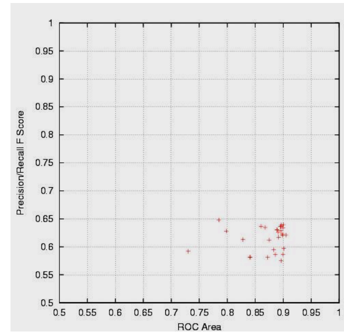


Problem 2

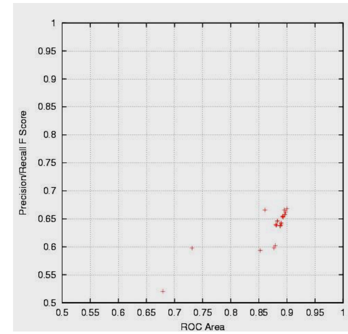


37

Problem 1



Problem 2



38

Many Other Metrics

- Mitre F-Score
- Kappa score
- Balanced Accuracy
- RMSE (squared error)
- Log-loss (cross entropy)
- Calibration
  - reliability diagrams and summary scores
- …

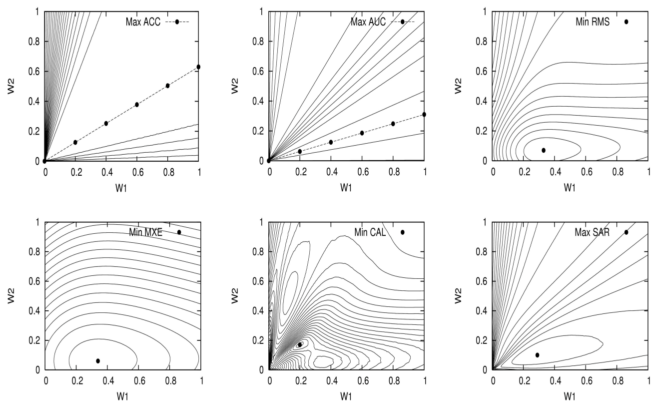
39

Summary

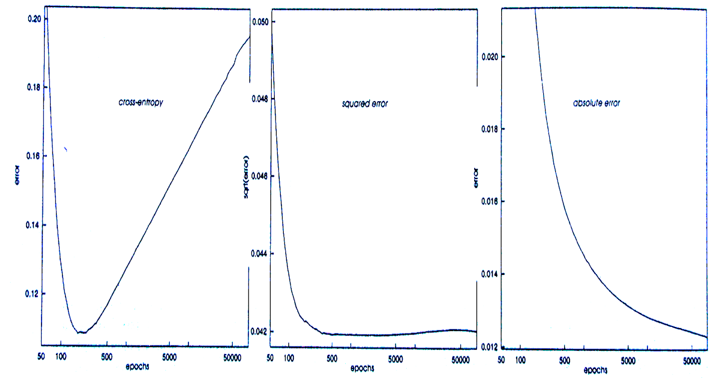
- the measure you optimize to makes a difference
- the measure you report makes a difference
- use measure appropriate for problem/community
- accuracy often is not sufficient/appropriate
- ROC is gaining popularity in the ML community
- only a few of these (e.g. accuracy) generalize easily to >2 classes

40

Different Models Best on Different Metrics

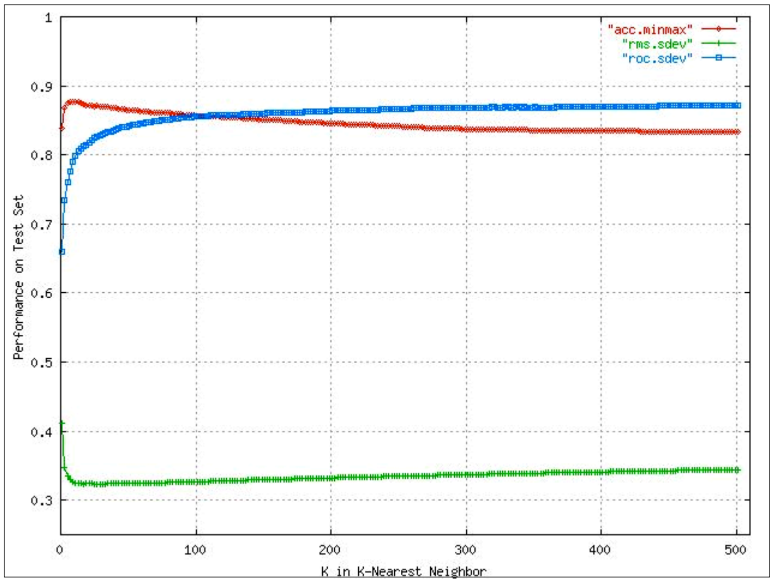


41



[Andreas Weigend, Connectionist Models Summer School, 1993]

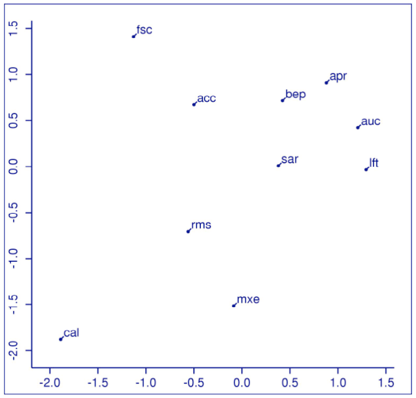
Figure 1. Out-of-sample errors as a function of training time for three error measures: cross-entropy, squared error, and absolute error. These three curves are from one and the same network that was trained with cross-entropy and tested (on the same data in each case) with different error measures.



Really does matter what you optimize!

44

2-D Multi-Dimensional Scaling



45