CSCE 478/878 Lecture 5: Evaluating Hypotheses

Stephen D. Scott
(Adapted from Tom Mitchell's slides)

October 5, 2006

Outline

- Sample error vs. true error
- Confidence intervals for observed hypothesis error
- Estimators
- Binomial distribution, Normal distribution, Central Limit Theorem
- Paired t tests
- Comparing learning methods
- ROC analysis

Two Definitions of Error

- The true error of hypothesis h with respect to target function f and distribution $\mathcal{D}$ is the probability that h will misclassify an instance drawn at random according to $\mathcal{D}$.

$$error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[f(x) \neq h(x)]$$

- The sample error of h with respect to target function f and data sample S ($|S| = n$) is the proportion of examples h misclassifies

$$error_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x) \neq h(x)),$$

where $\delta(f(x) \neq h(x))$ is 1 if $f(x) \neq h(x)$, and 0 otherwise.

- How well does $error_S(h)$ estimate $error_{\mathcal{D}}(h)$?

Problems Estimating Error

- Bias: If S is training set, $error_S(h)$ is optimistically biased

$$bias \equiv E[error_S(h)] - error_{\mathcal{D}}(h)$$

  For unbiased estimate ($bias = 0$), h and S must be chosen independently $\Rightarrow$ Don't test on training set!

  Don't confuse with inductive bias!

- Variance: Even with unbiased S, $error_S(h)$ may still *vary* from $error_{\mathcal{D}}(h)$

Estimators

Experiment:

1. Choose sample S of size n according to distribution $\mathcal{D}$

2. Measure $error_S(h)$

$error_S(h)$ is a random variable (i.e., result of an experiment)

$error_S(h)$ is an unbiased estimator for $error_{\mathcal{D}}(h)$

Given observed $error_S(h)$, what can we conclude about $error_{\mathcal{D}}(h)$?

Confidence Intervals

If

- S contains n examples, drawn independently of h and each other
- $n \geq 30$

Then

- With approximately 95% probability, $error_{\mathcal{D}}(h)$ lies in interval

$$error_S(h) \pm 1.96\sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

E.g. hypothesis h misclassifies 12 of the 40 examples in test set S:

$$error_S(h) = \frac{12}{40} = 0.30$$

Then with approx. 95% confidence,
$error_{\mathcal{D}}(h) \in [0.158, 0.442]$

Confidence Intervals

(cont'd)

If

- S contains n examples, drawn independently of h and each other
- $n \geq 30$

Then

- With approximately N% probability, $error_{\mathcal{D}}(h)$ lies in interval

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

where

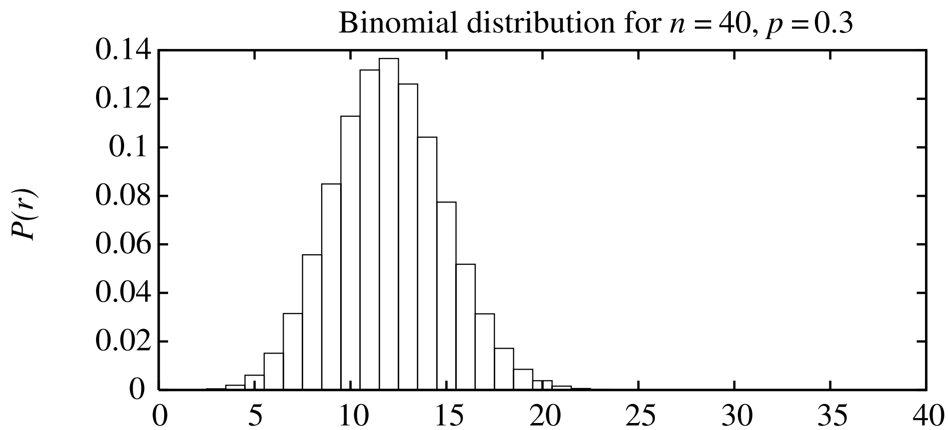
N%:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

Why?

$error_S(h)$ **is a Random Variable**

Repeatedly run the experiment, each with different randomly drawn S (each of size n)

Probability of observing r misclassified examples:

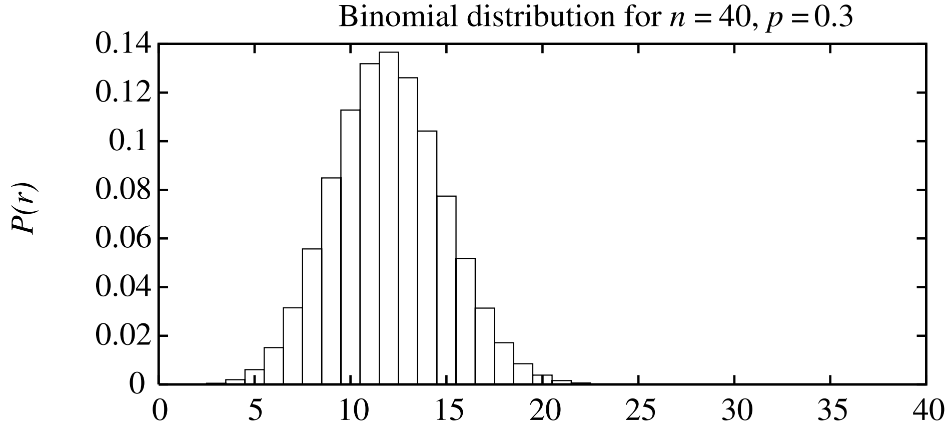


$$P(r) = \binom{n}{r} error_{\mathcal{D}}(h)^r (1 - error_{\mathcal{D}}(h))^{n-r}$$

I.e. let $error_{\mathcal{D}}(h)$ be probability of heads in biased coin, the $P(r) =$ prob. of getting r heads out of n flips

What kind of distribution is this?

Binomial Probability Distribution



$$P(r) = \binom{n}{r} p^r (1-p)^{n-r} = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

Probability $P(r)$ of r heads in n coin flips, if $p = \Pr(heads)$

- Expected, or mean value of X, $E[X]$ (= # heads on n flips = # mistakes on n test exs), is

$$E[X] \equiv \sum_{i=0}^{n} iP(i) = np = n \cdot error_{\mathcal{D}}(h)$$

- Variance of X is

$$Var(X) \equiv E[(X - E[X])^2] = np(1-p)$$

- Standard deviation of X, σ_X, is

$$\sigma_X \equiv \sqrt{E[(X - E[X])^2]} = \sqrt{np(1-p)}$$

Approximate Binomial Dist. with Normal

$error_S(h) = r/n$ is binomially distributed, with

- mean $\mu_{error_S(h)} = error_{\mathcal{D}}(h)$ (i.e. unbiased est.)
- standard deviation $\sigma_{error_S(h)}$

$$\sigma_{error_S(h)} = \sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

(i.e. increasing n decreases variance)

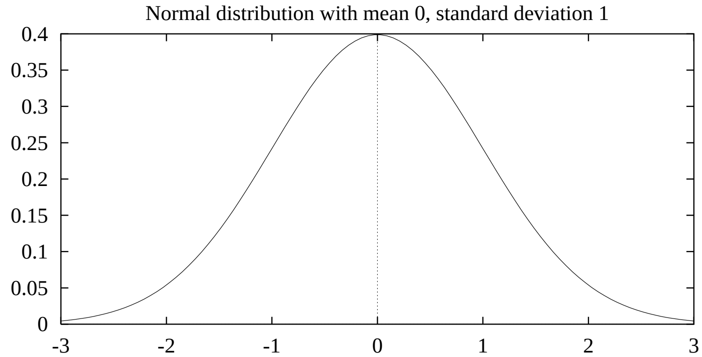
Want to compute confidence interval = interval centered at $error_{\mathcal{D}}(h)$ containing N% of the weight under the distribution (difficult for binomial)

Approximate binomial by normal (Gaussian) dist:

- mean $\mu_{error_S(h)} = error_{\mathcal{D}}(h)$
- standard deviation $\sigma_{error_S(h)}$

$$\sigma_{error_S(h)} \approx \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

Normal Probability Distribution



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

- Defined completely by μ and σ
- The probability that X will fall into the interval (a, b) is given by

$$\int_a^b p(x)dx$$

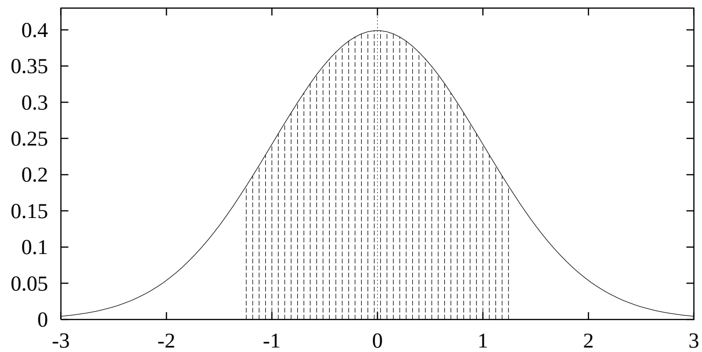
- Expected, or mean value of X, $E[X]$, is

$$E[X] = \mu$$

- Variance of X is $Var(X) = \sigma^2$
- Standard deviation of X, σ_X, is

$$\sigma_X = \sigma$$

Normal Probability Distribution
(cont'd)

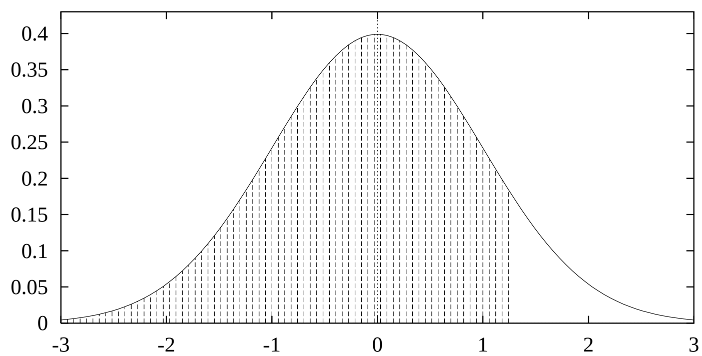


80% of area (probability) lies in $\mu \pm 1.28\sigma$

N% of area (probability) lies in $\mu \pm z_N\,\sigma$

N%:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

Can also have one-sided bounds:



N% of area lies $< \mu + z'_N\,\sigma$ or $> \mu - z'_N\sigma$, where $z'_N = z_{100-(100-N)/2}$

N%:	50%	68%	80%	90%	95%	98%	99%
z'_N:	0.0	0.47	0.84	1.28	1.64	2.05	2.33

Confidence Intervals Revisited

If

- S contains n examples, drawn independently of h and each other
- $n \geq 30$

Then

- With approximately 95% probability, $error_S(h)$ lies in interval

$$error_{\mathcal{D}}(h) \pm 1.96\sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

Equivalently, $error_{\mathcal{D}}(h)$ lies in interval

$$error_S(h) \pm 1.96\sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

which is approximately

$$error_S(h) \pm 1.96\sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

(One-sided bounds yield upper or lower error bounds)

Central Limit Theorem

How can we justify approximation?

Consider a set of independent, identically distributed random variables $Y_1 \ldots Y_n$, all governed by an arbitrary probability distribution with mean μ and finite variance σ^2. Define the sample mean

$$\bar{Y} \equiv \frac{1}{n} \sum_{i=1}^{n} Y_i$$

Note that $\bar{Y}$ is itself a random variable, i.e. the result of an experiment (e.g. $error_S(h) = r/n$)

Central Limit Theorem: As $n \rightarrow \infty$, the distribution governing $\bar{Y}$ approaches a Normal distribution, with mean μ and variance σ^2/n

Thus the distribution of $error_S(h)$ is approximately normal for large n, and its expected value is $error_{\mathcal{D}}(h)$

(Rule of thumb: $n \geq 30$ when estimator's distribution is binomial, might need to be larger for other distributions)

Calculating Confidence Intervals

1. Pick parameter p to estimate
   - $error_{\mathcal{D}}(h)$
2. Choose an estimator
   - $error_S(h)$
3. Determine probability distribution that governs estimator
   - $error_S(h)$ governed by binomial distribution, approximated by normal when $n \geq 30$
4. Find interval (L, U) such that N% of probability mass falls in the interval
   - Could have $L = -\infty$ or $U = \infty$
   - Use table of z_N or z'_N values (if distrib. normal)

Difference Between Hypotheses

Test h_1 on sample S_1, test h_2 on S_2, $S_1 \cap S_2 = \emptyset$

1. Pick parameter to estimate

$$d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$$

2. Choose an estimator

$$\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$$

(unbiased)

3. Determine probability distribution that governs estimator (difference between two normals is also normal, variances add)

$$\sigma_{\hat{d}} \approx \sqrt{\frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_{S_2}(h_2))}{n_2}}$$

4. Find interval (L, U) such that N% of prob. mass falls in the interval: $\hat{d} \pm z_n \, \sigma_{\hat{d}}$

(Can also use $S = S_1 \cup S_2$ to test h_1 and h_2, but not as accurate; interval overly conservative)

Paired t test to compare h_A,h_B

1. Partition data into k disjoint test sets $T_1, T_2, \ldots, T_k$ of equal size, where this size is at least 30
2. For i from 1 to k, do

$$\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$$

3. Return the value $\bar{\delta}$, where

$$\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i$$

N% confidence interval estimate for d:

$$\bar{\delta} \pm t_{N,k-1}\ s_{\bar{\delta}}$$

$$s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} \left(\delta_i - \bar{\delta}\right)^2}$$

t plays role of z, s plays role of σ

t test gives more accurate results since std. deviation approximated and test sets for h_A and h_B not independent

Comparing Learning Algorithms L_A and L_B

What we'd like to estimate:

$$E_{S \subset \mathcal{D}}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$$

where $L(S)$ is the hypothesis output by learner L using training set S

I.e., the expected difference in true error between hypotheses output by learners L_A and L_B, when trained using randomly selected training sets S drawn according to distribution $\mathcal{D}$

But, given limited data D_0, what is a good estimator?

- Could partition D_0 into training set S_0 and testing set T_0, and measure

$$error_{T_0}(L_A(S_0)) - error_{T_0}(L_B(S_0))$$

- Even better, repeat this many times and average the results (next slide)

Comparing learning algorithms L_A and L_B (cont'd)

k-fold Cross Validation

1. Partition data D_0 into k disjoint test sets $T_1, T_2, \ldots, T_k$ of equal size, where this size is at least 30
2. For i from 1 to k, do

   (use T_i for the test set, and the remaining data for training set S_i)

   - $S_i \leftarrow D_0 - T_i$
   - $h_A \leftarrow L_A(S_i)$
   - $h_B \leftarrow L_B(S_i)$
   - $\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$

3. Return the value $\bar{\delta}$, where

$$\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i$$

Comparing learning algorithms L_A and L_B (cont'd)

- Notice we'd like to use the paired t test on $\bar{\delta}$ to obtain a confidence interval

- Not really correct, because the training sets in this algorithm are not independent (they overlap!)

- More correct to view algorithm as producing an estimate of

$$E_{S \subset D_0}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$$

  instead of

$$E_{S \subset \mathcal{D}}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$$

- But even this approximation is better than nothing

ROC Analysis

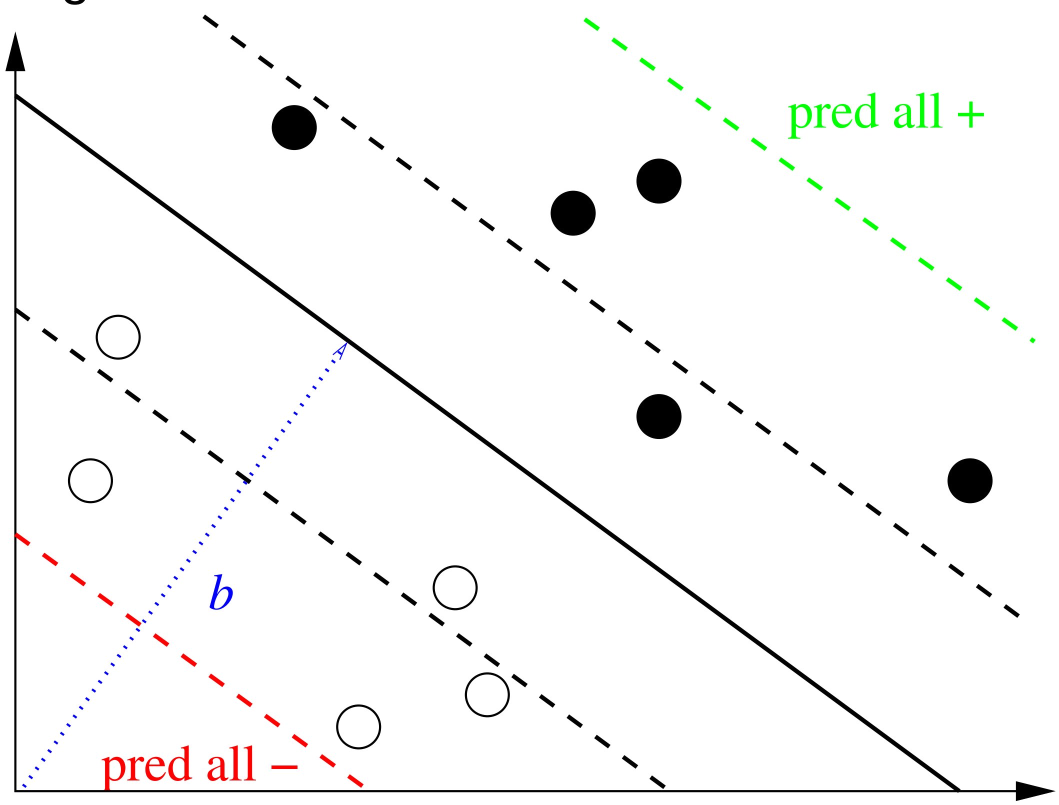
- So far, we've looked at a single error rate to compare hypotheses/learning algorithms/etc.

- This may not tell the whole story:
  - 1000 test examples: 20 positive, 980 negative
  - h_A gets 2/20 pos correct, 965/980 neg correct, for accuracy of $(2 + 965)/(20 + 980) = 0.967$
  - Pretty impressive, except that always predicting negative yields accuracy $= 0.980$
  - Would we rather have h_B, which gets 19/20 pos correct and 930/980 neg, for accuracy $= 0.949$?
  - Depends on how important the positives are, i.e. frequency in practice and/or cost (e.g. cancer diagnosis)

- Can separately report false positive (FP) and false negative (FN) error rates, but we can give even more detail than that

ROC Analysis
(cont'd)

- Consider an ANN or SVM
- Normally threshold at 0, but what if we changed it?
- Keeping weight vector constant while changing threshold = holding hyperplane's slope fixed while moving along its normal vector



- I.e. get a set of classifiers, one per labeling of test set

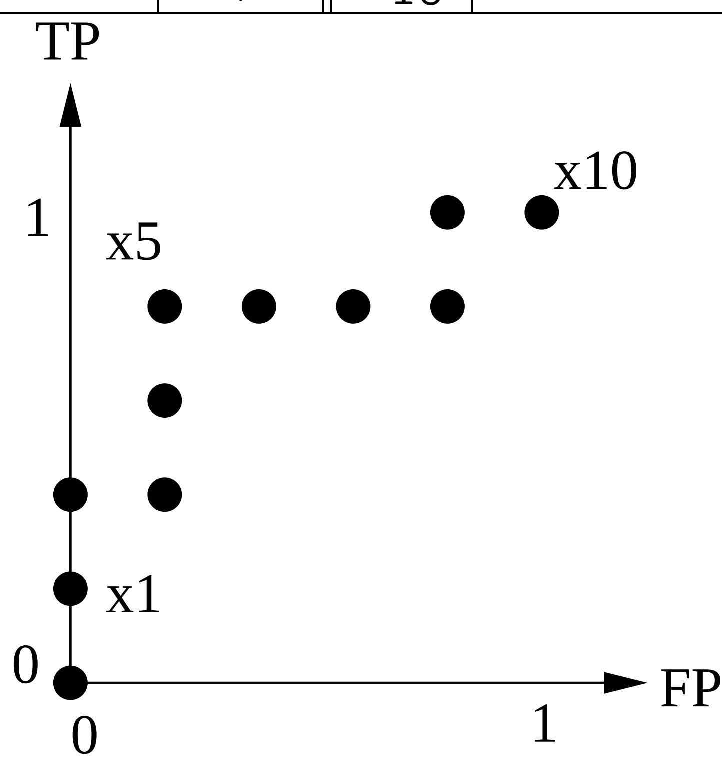
ROC Analysis

Plotting TP versus FP error

- Consider the "always $-$" hyp. What is its FP rate? Its TP rate? What about the "always $+$" hyp?

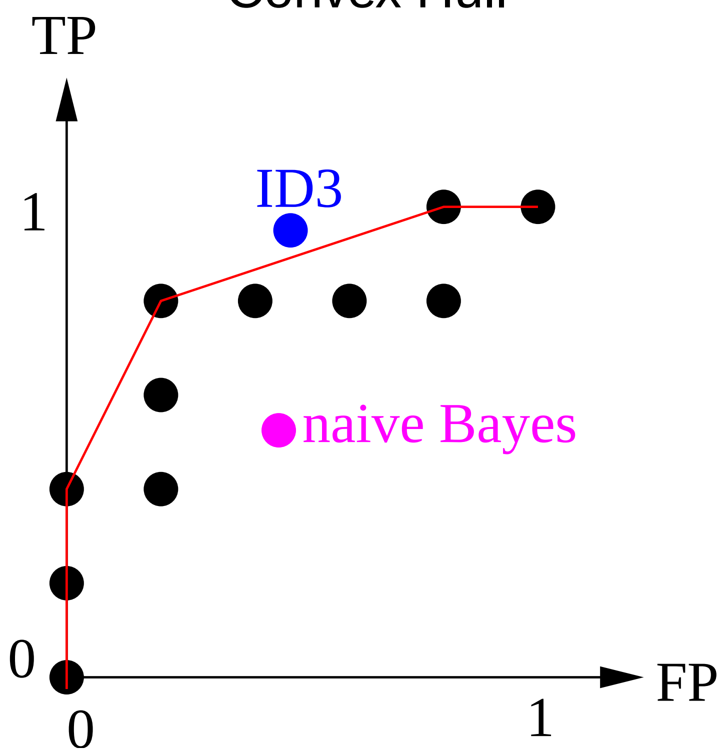
- In between the extremes, we plot TP versus FP by sorting the test examples by the SVM's weighted sums:

Ex	$\vec{w} \cdot \vec{x}$	label	Ex	$\vec{w} \cdot \vec{x}$	label
x_1	169.752	$+$	x_6	-12.640	$-$
x_2	109.200	$+$	x_7	-29.124	$-$
x_3	19.210	$-$	x_8	-83.222	$-$
x_4	1.905	$+$	x_9	-91.554	$+$
x_5	-2.75	$+$	x_{10}	-128.212	$-$



ROC Analysis

Convex Hull



- The convex hull of the ROC curve yields a collection of classifiers, each optimal under different conditions
  - If FP cost = FN cost, then draw a line with slope $|N|/|P|$ at $(0, 1)$ and drag it towards convex hull until you touch it; that's your operating point
  - Can use as a classifier any part of the hull since can randomly select between two classifiers
- Can also compare curves against "single-point" classifiers when no curves available
  - In plot, ID3 better than our SVM iff negatives scarce; nB never better

ROC Analysis

Miscellany

- What is the worst possible ROC curve?

- One metric for measuring a curve's goodness: area under curve (AUC):

$$\frac{\sum_{x_+ \in P} \sum_{x_- \in N} I(h(x_+) > h(x_-))}{|P|\,|N|}$$

  i.e. rank all examples by confidence in "$+$" prediction, count the number of times a positively-labeled example (from P) is ranked above a negatively-labeled one (from N), then normalize

  - What is the best value?

  - Distribution approximately normal if $|P|, |N| > 10$, so can find confidence intervals

  - Catching on as a better scalar measure of performance than error rate

- ROC analysis possible (though tricky) with multi-class problems

ROC Analysis

Miscellany (cont'd)

- Can use ROC curve to modify classifiers, e.g. re-label decision trees
- What does "ROC" stand for?
  - "Receiver Operating Characteristic" from signal detection theory, where binary signals are corrupted by noise
  - Use plots to determine how to set threshold to determine presence of signal
  - Threshold too high: miss true hits (TP rate low), too low: too many false alarms (FP rate high)
- Alternatives to ROC: cost curves and precision-recall curves

Topic summary due in 1 week!