Machine Learning

Topic: Evaluating Hypotheses

How do you tell something is better?

Assume we have an error measure….

- How do we tell if it measures something useful?

  To measure intelligence, which is a better? {grades, IQ, salary}

- If it is useful, how precise/unbiased/noisy is it?

- How much of a difference in the measure is required to say things two things are truly "different"?

  Maria's IQ is 103. Bob's is 101. Does that make her "smarter"?

What's a useful measure for a…

- Classifier (Decision tree)

  An idea: Count how often the classifier is wrong

- Regressor (Linear regression)

  An idea: the distance between predicted values and observed values

- Probability Mass (or density) Estimator

  Pick the distribution that maximizes the likelihood of the data?

  Pick the distribution that "looks" the most "reasonable"?

- Ranker (like a search engine) : Rank order?

- User interface widget: User satisfaction?

Definitions of Error

- $error_D(h)$ is the *true error* of hypothesis ***h*** with respect to the target function ***f*** and data distribution ***D***. It is the probability ***h*** will misclassify an instance drawn at random according to ***D***.

- $error_s(h)$ is the *sample error* of hypothesis ***h*** with respect to the target function ***f*** and data sample set ***S***. It is the proportion of examples in ***S*** that ***h*** misclassifies.

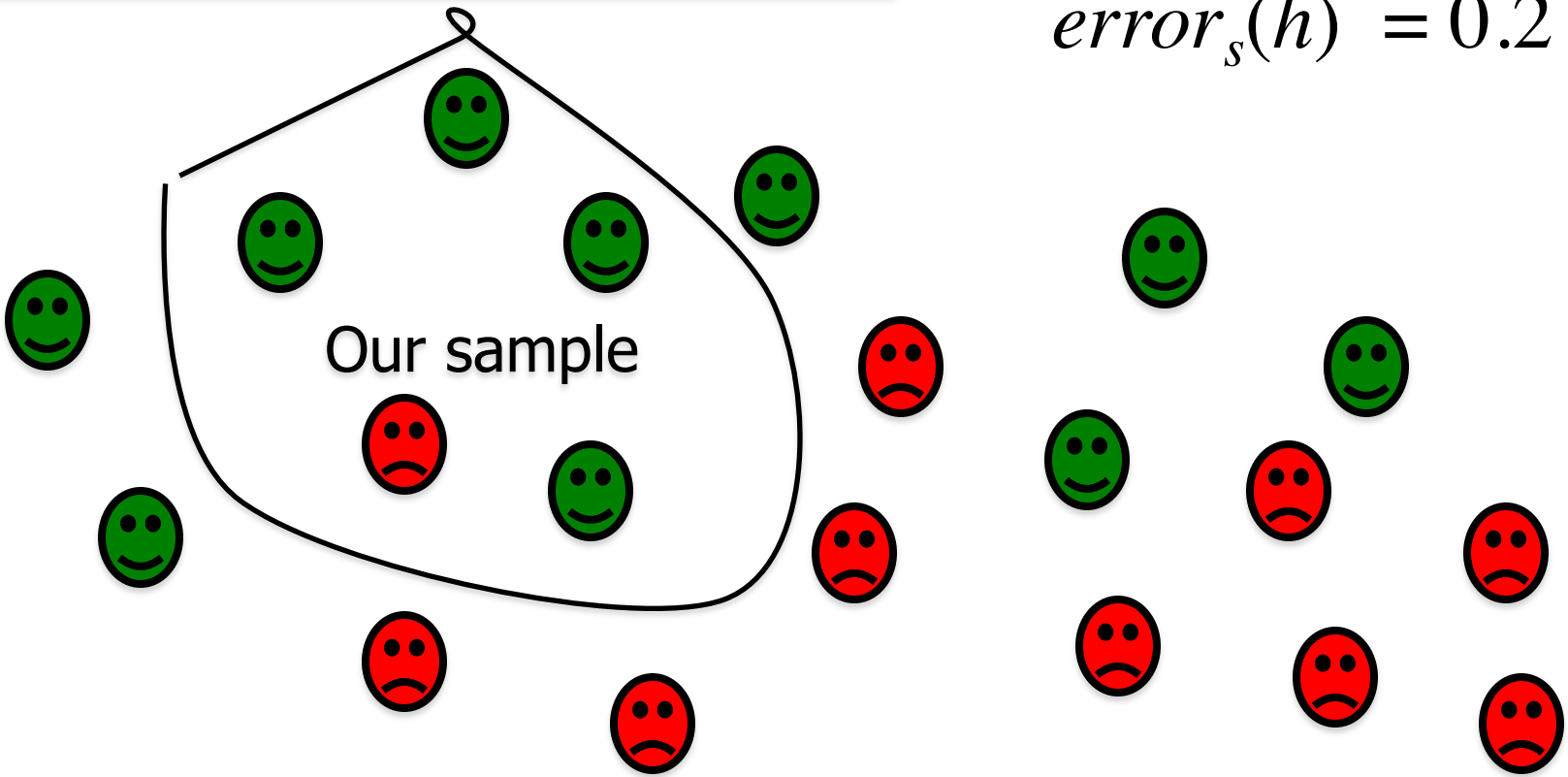
True Error vs Sample Error

Misclassified: h(x) != f(x)

Correctly classified: h(x) = f(x)

$$error_D(h) = 0.5$$

$$error_s(h) = 0.2$$



Sample Error: It's all we have

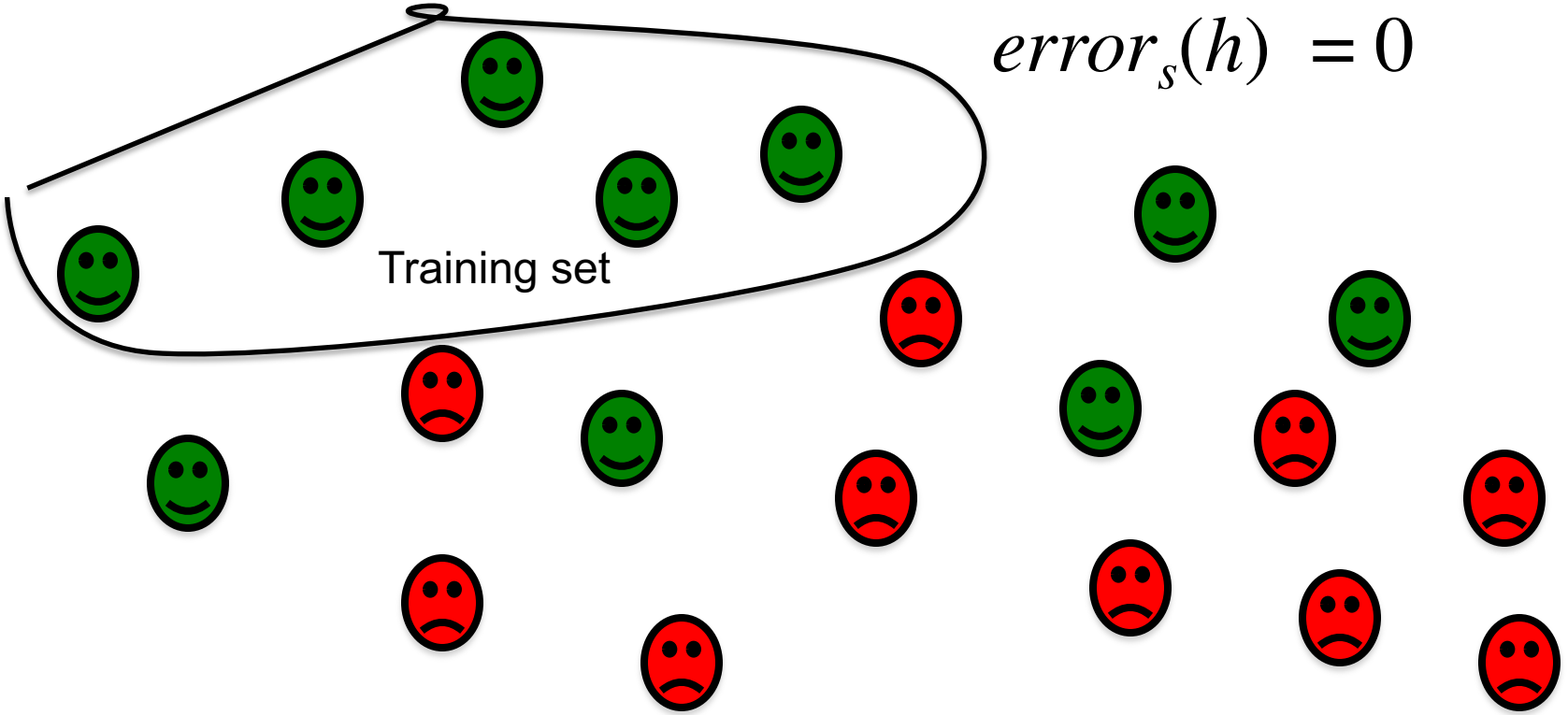
Generally, we never know the true error $error_D(h)$.
We only get to see the sample error $error_S(h)$.

How well does the sample error estimate the true error?

Can we set conditions for our experiment so that we can get an estimate that is good enough for our needs?

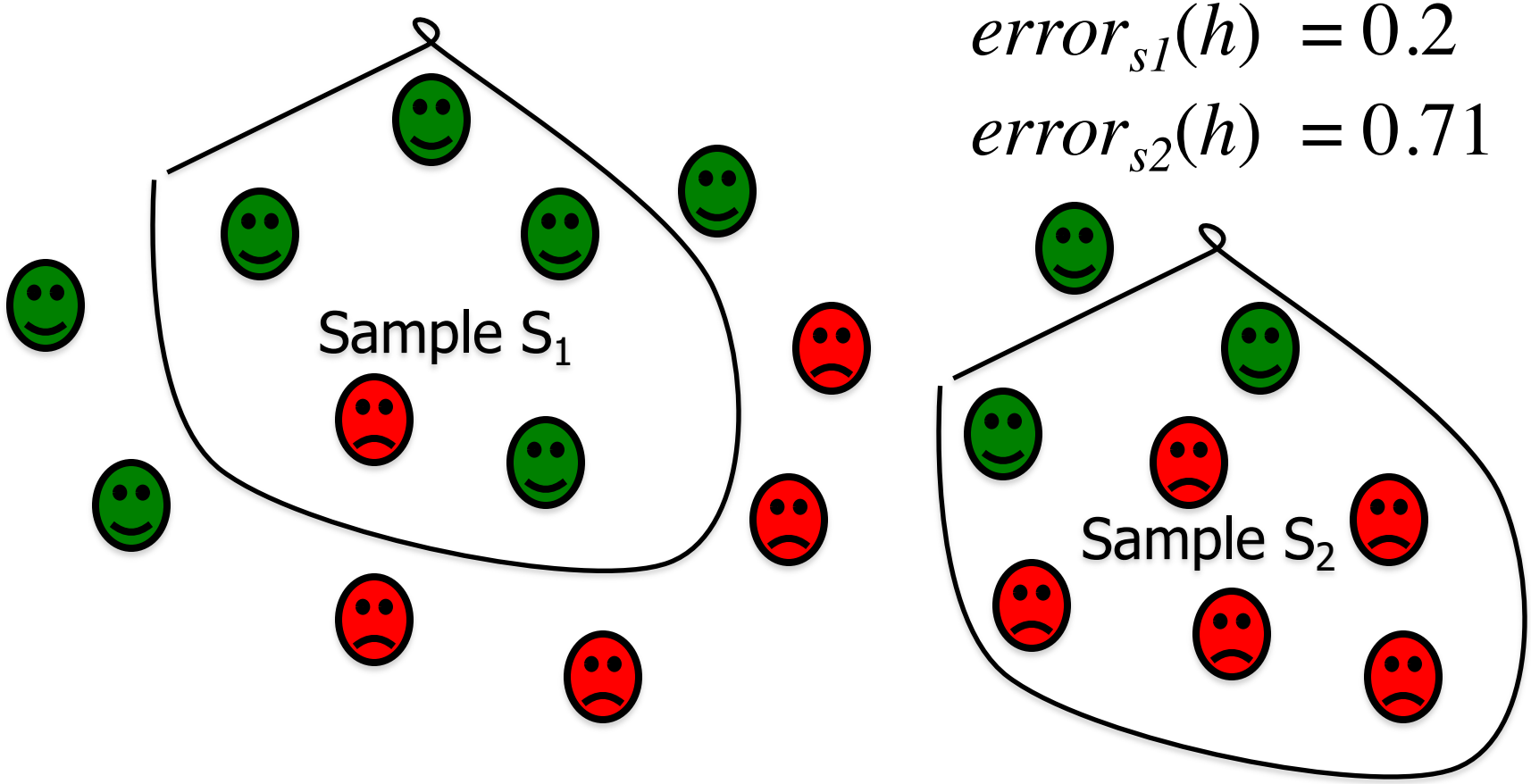
Problems Estimating Error

- BIAS: If S is the training set, $error_s(h)$ is optimistically biased. For an unbiased estimate we need a validation set that was not used in training.



Problems Estimating Error

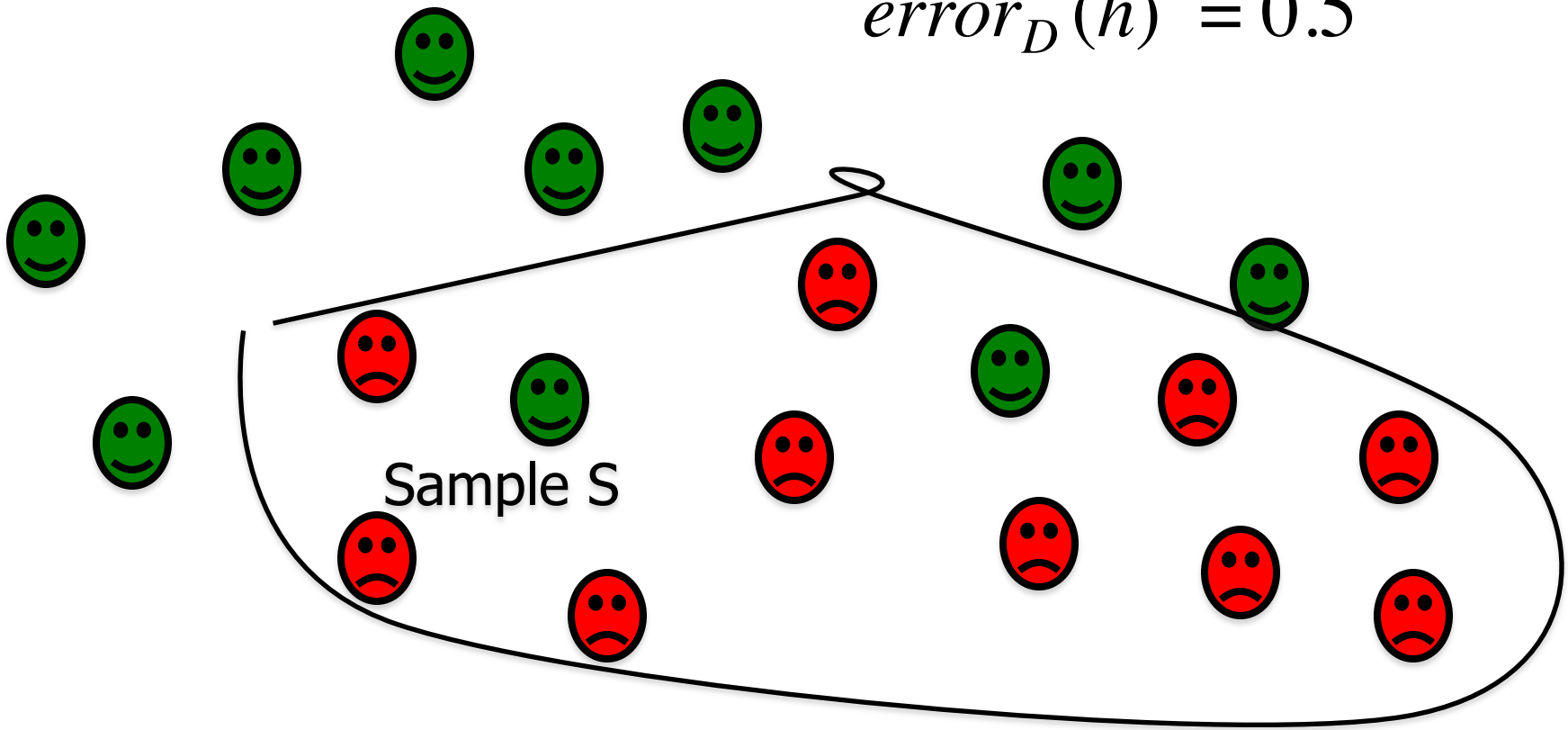
- Variance: Even without bias, $error_s(\boldsymbol{h})$ may still vary from $error_D(\boldsymbol{h})$



Q: Why not just take one bigger sample?

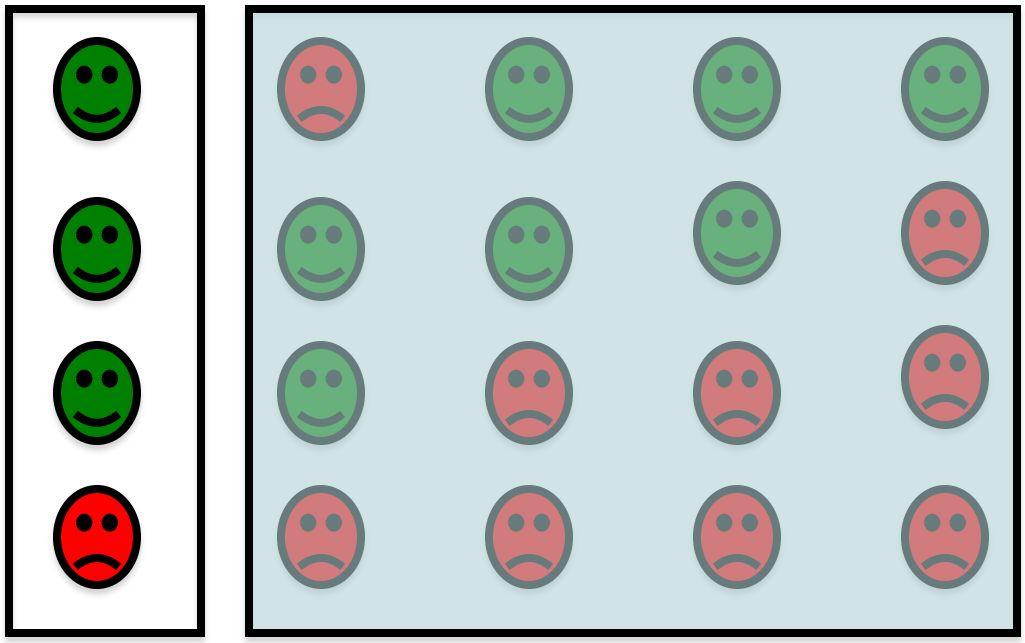
A: From one sample mean, you can't tell how $error_s(\boldsymbol{h})$ varies from $error_D(\boldsymbol{h})$

$$error_S(h) = 0.83$$
$$error_D(h) = 0.5$$



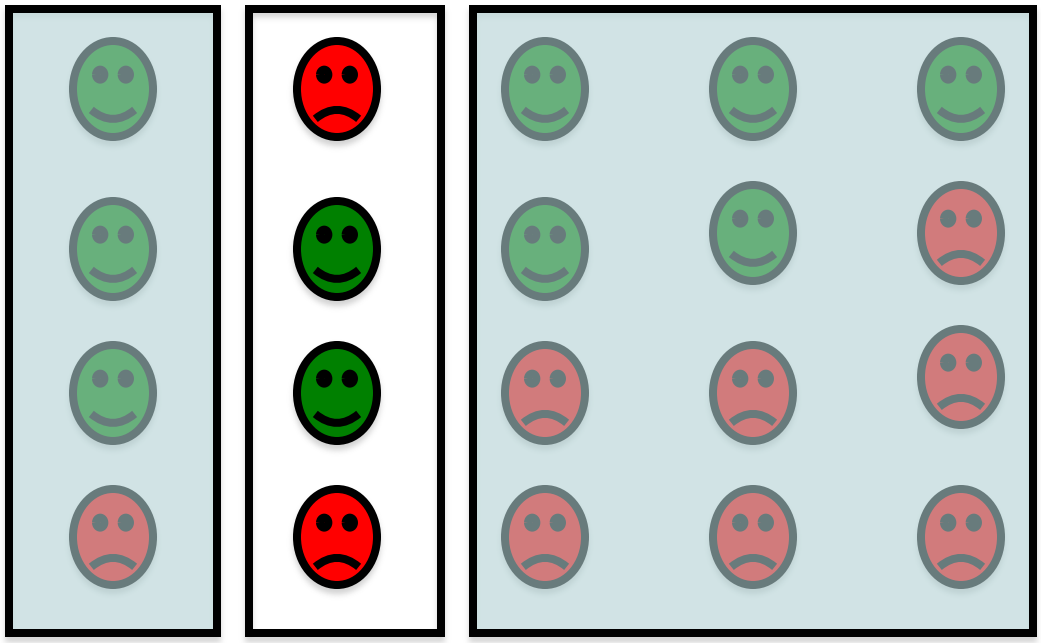
N-fold cross validation

- Spilt data into N groups.
- Train on N-1 groups.
- Validate on the Nth.
- Rotate, repeat.



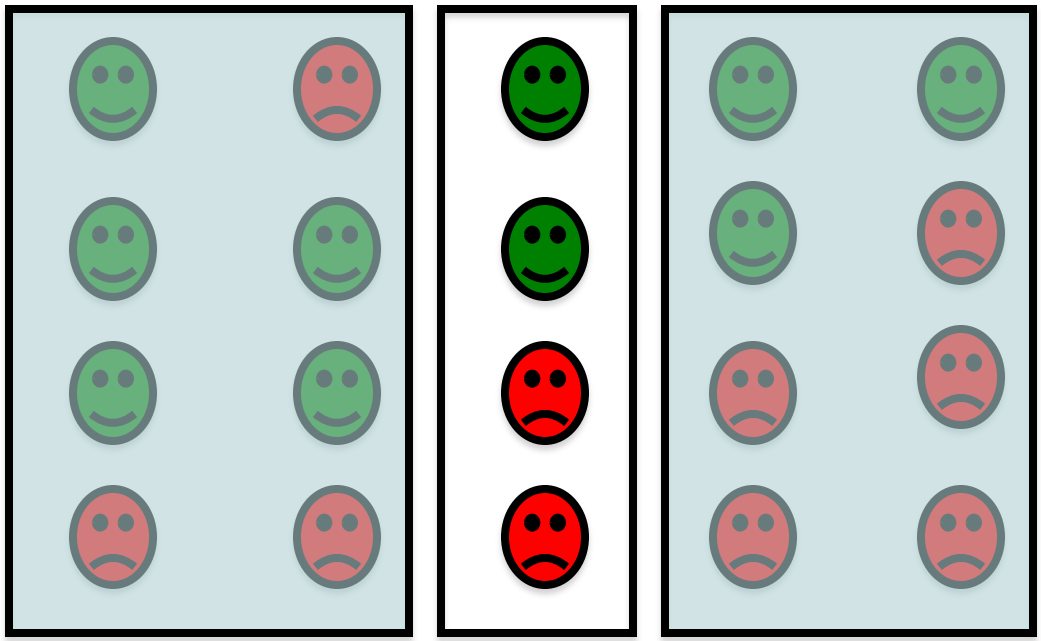
N-fold cross validation

- Spilt data into N groups.
- Train on N-1 groups.
- Validate on the Nth.
- Rotate, repeat.



N-fold cross validation

- Spilt data into N groups.
- Train on N-1 groups.
- Validate on the Nth.
- Rotate, repeat.



Precision vs Recall

Classifiers are often evaluated with an eye towards their being search engines. (e.g. labeling documents as either relevant or not to a search query). In this case people often use the following measures:

$$precision \qquad p = \frac{tp}{tp + fp}$$

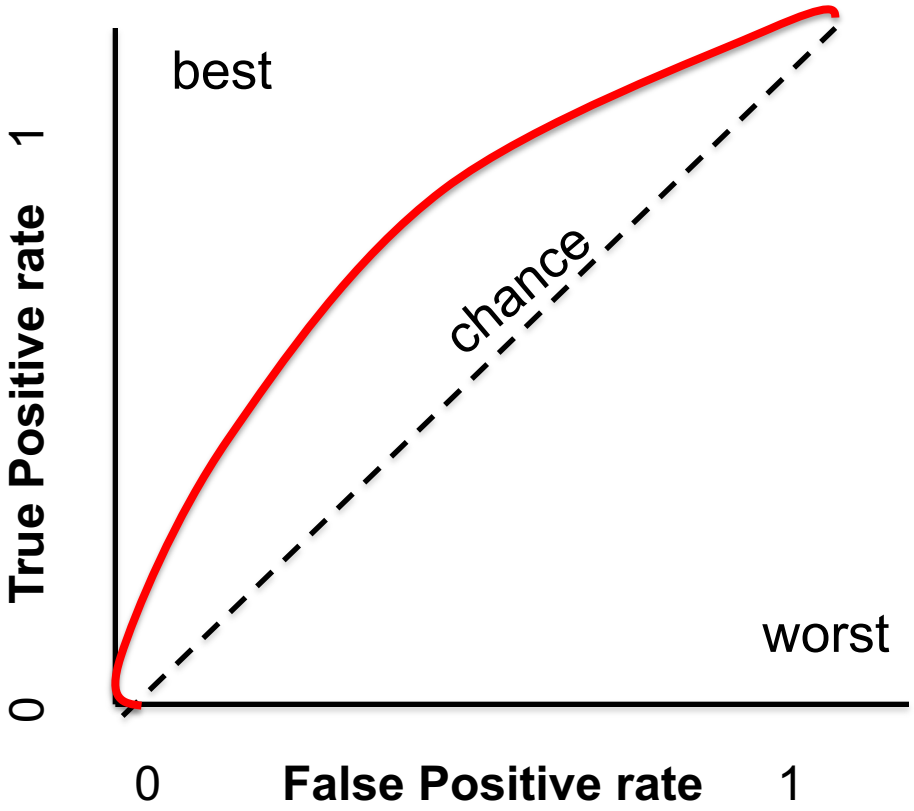
$$recall \qquad r = \frac{tp}{tp + fn}$$

$$F - measure \qquad F = 2\frac{p \cdot r}{p + r}$$

Machine's Classification	True Classification: True	True Classification: False
True	True positive (tp)	False positive (fp)
False	False negative (fn)	True negative (tn)

ROC

The Receiver Operating Characteristic (AKA ROC Curve) curve shows the tradeoff as you adjust parameters of your system.



Confusion Matrix

- Lets us see which things the classifier is mixing up. Helps direct improvement.

Correct Classification

Machine's Classification	Dog	Coyote	Cactus	Road Runner
Dog	8	5	0	2
Coyote	2	5	0	2
Cactus	0	0	8	2
Road Runner	0	0	2	4

Experiment

1. Choose sample ***S*** of size ***n*** using distribution ***D***
2. Measure $error_s(h)$

Question: What can we conclude about $error_D(h)$ from $error_s(h)$?

Answer: That's what we're here to learn today.

Coin flips

- Assume an unbiased coin X that takes two values $\{0,1\}$.
- Let K be the number we get if we flip the coin n times and add up the values of all the flips.
- What is the expected value of K?
- Assume $n = 5$

  How likely is K to be 0?

  How likely is K to be $n/2$?
- What distribution models this?

Some definitions

- A **Bernoulli Trial** is an experiment whose outcome is random (and has one of two outcomes (e.g. heads or tails). Think of it is a Boolean random variable, ***X.***

- A set of random $\{X_1, X_2, \ldots X_n\}$variables is **independent and identically distributed (IID)** if all variables in the set are mutually independent and all are governed by the same probability distribution ***D***.

Back to the coin example…

- If all coin flips use the same coin, we assume that they are IID Bernoulli trials
- This is modeled by the Binomial Distribution

$$P(K=k) = B(n,k,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Here, K is the summed value of the n coin flips and p is the probability of heads.

$$\text{(recall this->)} \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

The Normal Distribution

- As ***n*** goes to infinity, the Normal distribution approximates the Binomial distribution, if you set the standard deviation σ and mean μ correctly.

$$P(K = k) = B(n, k, p) \approx N(\mu, \sigma^2)$$

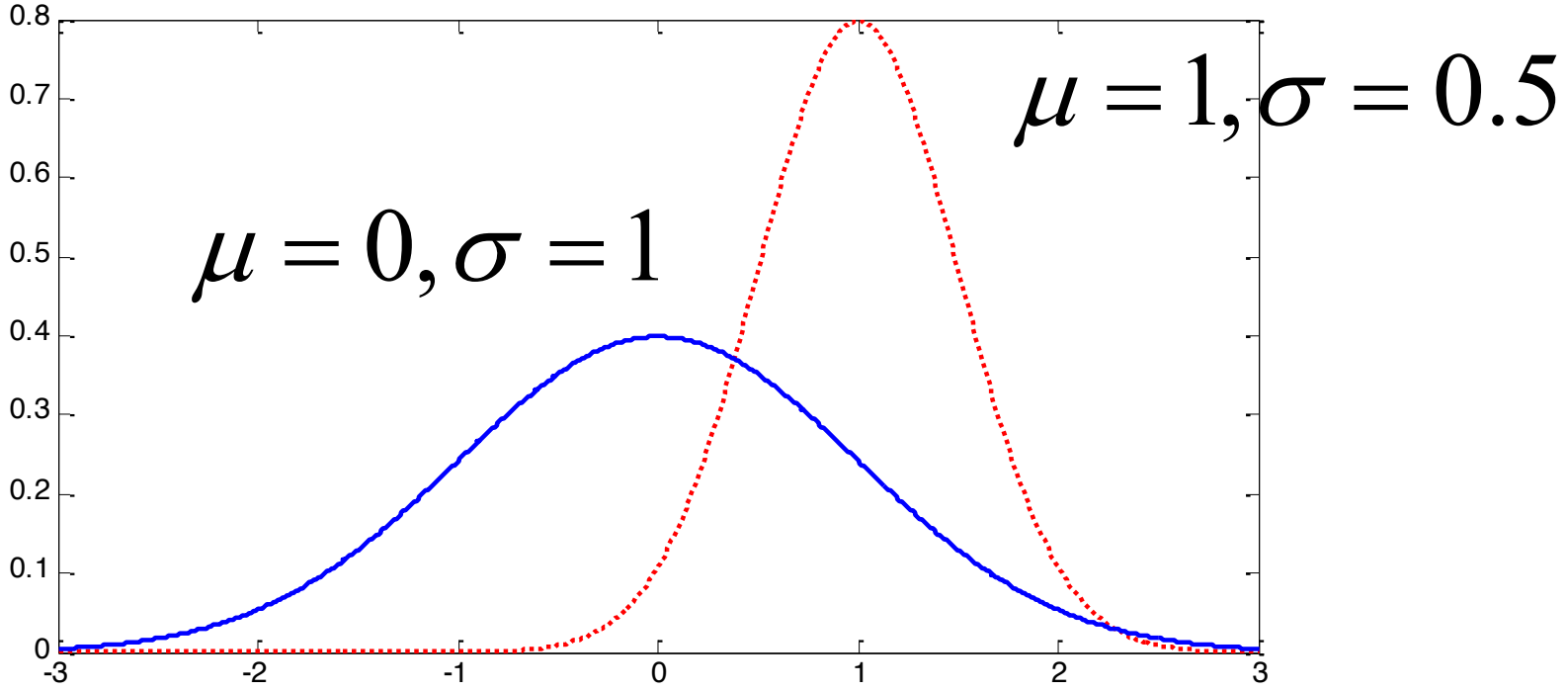
$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The Normal (Gaussian) Distribution

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

mean

variance



Central Limit Theorem

- Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample of size ***n***…i.e. a set of IID discrete random variables from some distribution ***D*** with expected value μ and variance σ^2.
- Define the sample average $\bar{x}$ as…

$$\bar{x} = \frac{1}{n} K = \frac{1}{n} \sum_{1}^{n} X_n$$

- For large n, the distribution of $\bar{x}$ is approximated by the normal distribution.

- **Important:** The distribution for the sample average approaches normality regardless of the shape of the distribution D governing our random samples X_i.

Why the previous slides matter

- Classification is like a coin flip, you're either right or wrong.
- If classification is independent, then the number of correct classifications K is governed by a Binomial distribution
- If the Binomial distribution is approximated by the Normal distribution we can use what we know about the Normal distribution.
- The Normal distribution lets us estimate how close the TRUE error is to the SAMPLE error.

How many samples do I need…

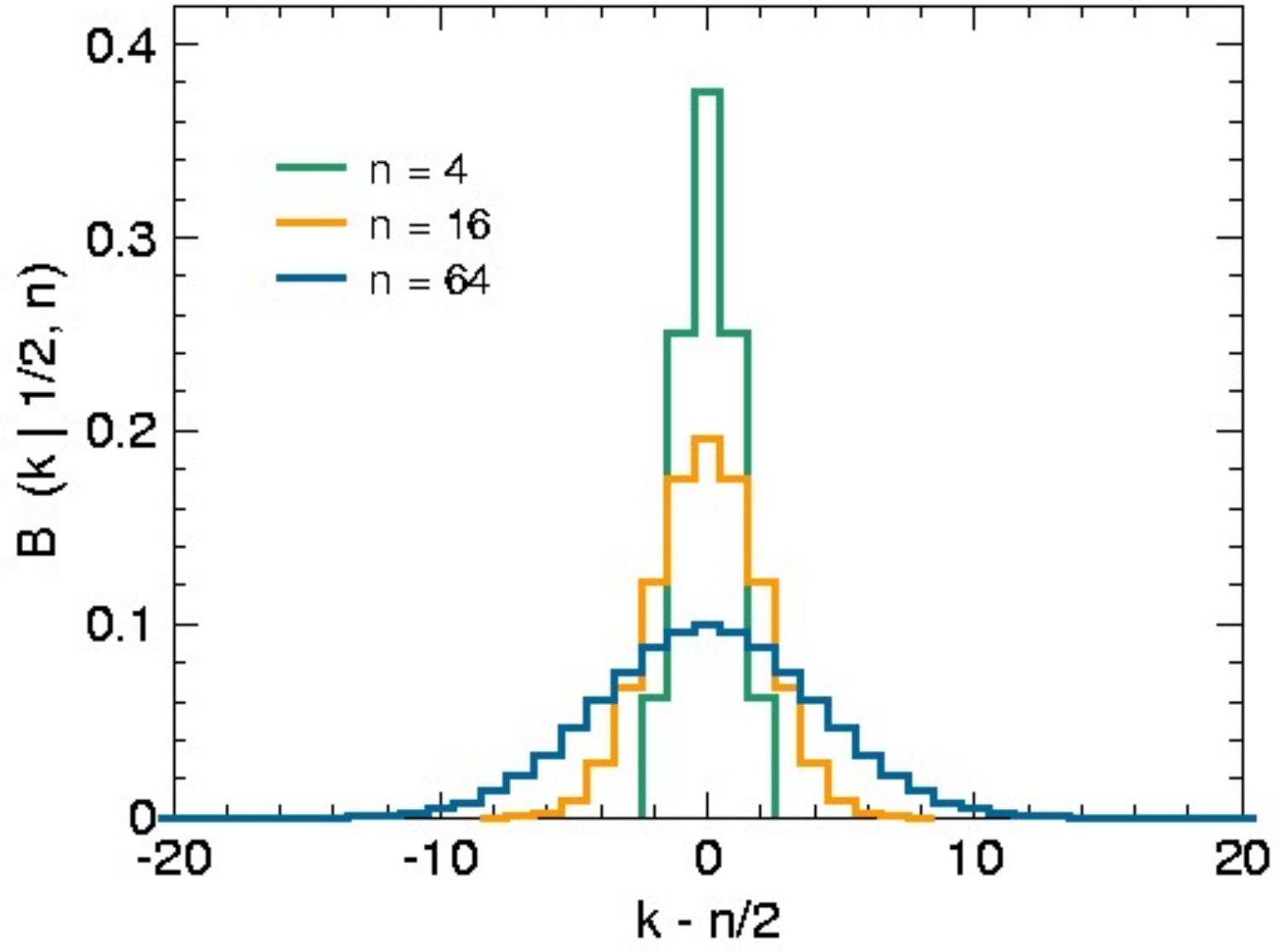
…before my sample's distribution is approximately normal?

More is always better. The more samples you have, the closer it gets to a normal distribution.

Rule of thumb: have at least 30 IID trials.

(let's look)

The binomial distribution as *n* grows



Source: http://en.wikipedia.org/wiki/File:Bindis-sym.png

Running a statistical test

1. Pick a parameter to estimate
2. Choose an estimator
3. Determine the probability distribution governing the estimator
4. Find the interval such that N% of the probability mass falls in that interval
5. The parameter has a N% chance of falling in that interval.

Confidence Intervals: Estimating a value

1. Pick a parameter to estimate

   $error_D(h)$

2. Choose an estimator

   $error_s(h)$

3. Determine the probability distribution governing the estimator

   $error_s(h)$ governed by Binomial, approximated by Normal when $n > 30$ (and bigger values for n are always better)

4. Find the interval such that N% of the probability mass falls in that interval

   Use your favorite statistics software, or look up z_N values.

How many samples do I need…

…to give me good confidence intervals (assuming we already have a normal distribution)?

The standard deviation of the sample mean is related to the standard deviation of the population σ and the size of the sample, n by the following:

$$SD_{\bar{X}} = \sigma / \sqrt{n}$$

Practical result: to decrease uncertainty in a mean estimate by a factor of n requires n^2 observations.

Setting 95% confidence interval size

- Recall that $SD_{\bar{X}} = \sigma / \sqrt{n}$
- For a normal distribution, 95% of the mass is within 2 standard deviations of the mean.
- For estimating a sample mean, an approximate 95% confidence interval has the form…

$$(\bar{x} - 2\sigma / \sqrt{n}, \bar{x} + 2\sigma / \sqrt{n})$$

- So, the 95% confidence interval width is

$$W = 4\sigma / \sqrt{n}$$

A rule of thumb

- If the sample *S* contains n IID examples drawn according to the distribution of *D*
- And $n >= 30$
- Then, the true error has a 95% chance of falling in the range…

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

- For a different % confidence, substitute the Z=1.96 value with the appropriate Z. See Chapter 5 of Machine Learning for more…

Student's t-test Facts

- One of the most commonly used statistical tests
- Assumes normally distributed data
- Different variants for different questions….

one sample t-test: Is a known population mean μ different from the mean of a sample population?

independent samples t-test: Are the means of two normally distributed populations equal?

paired samples t-test: Is 0 the mean difference between paired responses measured on the same data ?

Student's t-test Fact(oid)s

- The t-test was devised by William Gosset in 1908
- It was used to monitor the quality of Guinness Stout (beer).
- Gosset published the t-statistic under the name "student" because Guinness considered it a trade secret

one sample t-test

Abstract question: Is a known population mean μ different from the mean of a sample population?

- Example:

  We know μ =0.3 is the error rate ID3 has on categorizing a given data set.

  I trained 30 neural nets to categorize the same data set and the mean error rate was $\overline{x} = 0.2$

  Are neural nets better on this data set? Or was that a fluke?

- I'd use a one-sample t-test to find out.

one sample t-test

- Null Hypothesis: There is no significant difference between the sample mean and the population mean

Neural nets perform no better than ID3 on this data.

- Alternate Hypothesis: There is a significant difference between the sample mean and the population mean.

Neural nets DO perform better than ID3 on this data.

Paired samples t-test

Abstract question: Is 0 the mean difference between paired responses measured on the same data ?

- Example

  Does eating ice cream make you heavier?

  Take 1000 people.

  Weigh each of them.

  Feed each one an ice cream cone.

  Weigh each of them again.

- A paired-samples t-test is appropriate (Why?)

Paired samples t-test

- Null Hypothesis: There is no significant difference between the two sample means

Ice cream does not make you heavier

- Alternate Hypothesis: There is a significant difference between the two sample means

Ice cream makes you heavier. Or it makes you lighter. We didn't actually check which way the difference goes.

independent samples t-test

Abstract question: Are the means of two normally distributed populations equal?

- Example

  Is C4.5 better than ID3 on identifying "bad movies" from a database of 1000 labeled movies?

  I do the following 30 times:

    Train C4.5 on of 500 randomly-selected examples

    Test on the other 500

  I then repeat that for ID3

- An independent samples t-test is appropriate (Why?)

Common t-test pitfalls

- Data is not normally distributed (can't use a t-test)
- Not enough sample points (degrees of freedom)
- Using a paired-samples t-test on data where the samples aren't paired (use independent samples t-test, instead)
- Using a Student's independent samples t-test when the variances of the two sets are different (use Welch's t-test in this case)

Comparing populations: Which test?

Are the means of two populations equal?
What assumption does each test make?

Name of test	Samples IID	Gaussian	Paired	Both pops have same variance
Student's T-Test: Paired samples	X	X	X	X
Student's T-Test: Unpaired samples	X	X		X
Welch's T-test	X	X		
Wilcoxon signed-rank test	X		X	
Mann–Whitney *U* test	X			

Conclusions

- The error measure should captures what you really want to know….not what is easy to measure.
- Your measure may have variance/bias/noise. Therefore…
- Results are more meaningful when a statistical significance test is done.
- Many tests depend on the data being normally distributed
- By taking the sample average of a large set of IID trials, you can ensure normal-like data
- The t-test is a good, easy test to use…if you know when to use it and how