acc $\pm\, \epsilon$

CS 412

APR 14TH – CONCENTRATION BOUNDS

Administrivia

Midterm Graded by next Tuesday

Project meeting schedule + HW4 Out tonight

- Will reply as a comment on your gradescope submission
- HW4 Due next Thursday April 23rd, 11:30pm

Testing

Back to statistics

Suppose the ~~average~~ student carries \$20 in cash on average

- What is the probability a particular student carries \$100 in cash?

lowest probability

1 very wealthy student

everyone else has zero

$$P(X \geq 100) = \frac{1}{n}$$

$$k = \frac{1}{5} n$$

highest probability

→ k have \$100

n−k have \$0

maximize k

$$\frac{k \cdot 100 + (n-k) \cdot 0}{n} = \$20$$

Back to statistics

can't have negative cash

Suppose the average student carries $20 in cash

- What is the probability a particular student carries $100 in cash?

What is the mathematical notation for this problem?

- Define a random variable: let X be the number of dollars in a student's pocket
- So, what is $20? E[X]
- What are we trying to find? P(X>100)
- *Note: X must be non-negative (pretend debt isn't real)*

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$ *where f(x) is the pdf*

$$\mathrm{E}(X) = \int_0^a x f(x)\, dx + \int_a^\infty x f(x)\, dx \geq \int_a^\infty x f(x)\, dx \geq \int_a^\infty a f(x)\, dx = a \int_a^\infty f(x)\, dx = a \Pr(X \geq a)$$

non-negative function

- **Markov's Inequality:** $\Pr(X \geq a) \leq \mathrm{E}(X)/a$

Markov's Inequality

Markov's Inequality: $\Pr(X \geq a) \leq \mathrm{E}(X)/a$

This is a concentration bound, it shows us a bound on how the data is going to be concentrated

Suppose the average student carries \$20 in cash

- What is the probability a particular student carries \$100 in cash?

$$P(X \geq 100) \leq \frac{20}{100}$$

Markov's Inequality

E_{train}
E_{cv}
E_{test}
E_{out}

Markov's Inequality: $\Pr(X \geq a) \leq \mathrm{E}(X)/a$

This is a concentration bound, it shows us a bound on how the data is going to be concentrated

Suppose the average student carries \$20 in cash

- What is the probability a particular student carries \$100 in cash?
- 20/100 = 0.2

E_{out} – error of the model in the population

$= E_{test} \pm \epsilon$ wp δ

Markov's Inequality

Markov's Inequality: $\Pr(X \geq a) \leq \mathrm{E}(X)/a$

This is a concentration bound, it shows us a bound on how the data is going to be concentrated

Suppose the average student carries \$20 in cash

- What is the probability a particular student carries \$100 in cash?
- 20/100 = 0.2

Why is this going to be useful? What non-negative random variable do we care about?

- E_{out}! What is the actual error our model is going to have?:
- E_{cv} and E_{test} are just estimators. A lot of this math is also directly applicable to polling
- Are E_{cv} and E_{test} unbiased?

Not if we consider multiple models E_{out}

$-E_{test}$

Motivation

Suppose there was a measure on the ballot, and I'm trying to determine what proportion of the population (A) supports it.

- More critically, what am I interested in?

Motivation

Suppose there was a measure on the ballot, and I'm trying to determine what proportion of the population (A) supports it.

- More critically, what am I interested in? P(A>0.5)

Motivation

Suppose there was a measure on the ballot, and I'm trying to determine what proportion of the population (A) supports it.

- More critically, what am I interested in? P(A>0.5)
- How would I estimate this value? A poll? What are some problems?
  - How do I ask the question?
  - Who do I ask?
  - When do I ask?
  - How many people do I ask?

survey design

- Which of these are applicable to our problem of E_{out}?
  - Is my data representative of what I'm going to predict?
  - What if I had 15 polls? What should my reported value for A be?
  - Maybe take an average, but in ML, we want the **best** model, it makes sense that we should pay some penalty
  - This also explains why we only want to use the test data once – keeps the estimate unbiased

T

Motivation

$$E(E_{test}) = E(E_{out})$$

Markov's Inequality: $\Pr(X \geq a) \leq \mathrm{E}(X)/a$

What is the probability our actual error X=E_{out} is double our expected error E_{test}?

$$P(E_{out} \geq 2 \cdot E_{test}) \leq \frac{E(E_{out})}{2 \cdot E(E_{test})} = \frac{1}{2}$$

$E_{test} = 0.05$

$a \geq 2 \cdot E_{test}$

Motivation

Markov's Inequality: $\Pr(X \geq a) \leq \mathrm{E}(X)/a$

What is the probability our actual error X=E_{out} is double our expected error E_{test}?

- ½ -- that's not great
- What is our 95% confidence interval?

$$P(E_{out} > \varepsilon) < 0.05$$

$$E_{out} = E_{test} \pm E_{test}$$

w.p. 0.5

$$E_{test} = 0.05$$

$$\frac{E(X)}{a} = 0.05$$

$$a = \frac{E(X)}{0.05}$$

Motivation

Markov's Inequality: $\Pr(X \geq a) \leq \mathrm{E}(X)/a$

What is the probability our actual error X=E_{out} is double our expected error E_{test}?

- ½ -- that's not great
- What is our 95% confidence interval?
- $E_{out} = 20 * E_{test}$

Motivation

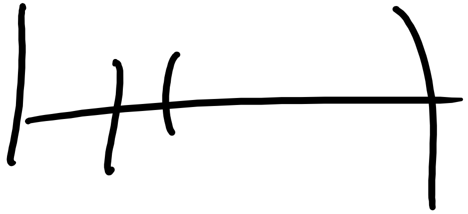
Markov's Inequality: $\Pr(X \geq a) \leq \mathrm{E}(X)/a$

What is the probability our actual error X=E_{out} is double our expected error E_{test}?

- ½ -- that's not great
- What is our 95% confidence interval?
- E_{out} = 20*E_{test} – ouch

Is this the best we can do?

- What else do we know about our error? Currently, what are we using?
- Only that it is non-negative! What else can we use?
- That we might know it's standard deviation, and that the error is bounded in [0,1]



Weaknesses of Markov's ineq

Markov's Inequality: $\Pr(X \geq a) \leq \mathrm{E}(X)/a$

Suppose LA gets an earthquake every 10 years, what is the probability that there will be an earth quake in the next 30 years?

Weaknesses of Markov's ineq

Markov's Inequality: $\Pr(X \geq a) \leq \mathrm{E}(X)/a$

$$\frac{E(x)}{30} = \frac{10}{30} = 1/3$$

Suppose LA gets an earthquake every ~~10~~ years, what is the probability that there will be an earth quake in the next 30 years?

$$P(X \leq 30) = 1 - P(X > 30) = 1 - 1/3 = 2/3$$

- 2/3

What is the probability that there will be an earthquake within the next 10 years?

X - length in years to the next earthquake

$$P(X \geq 10) \leq \frac{10 - E(x)}{10 - a}$$

$$E(x) = 10$$

$$P(X \geq 10) \leq 1$$

Markov's doesn't help

Weaknesses of Markov's ineq

How can we improve on the Markov bound?

Weaknesses of Markov's ineq

How can we improve on the Markov bound?

- Do we want to bound absolute error or relative error?

- What other information from our model are we not using?

Variance ~~stdev~~

data

Markov: $P(X \geq a) \leq \frac{E(X)}{a}$

$P(E_{out} > a)$

vs.

$P(|E_{out} - E_{test}| > \epsilon)$

Let $X = |E_{out} - E_{test}|$

X is now a non-negative random variable

Weaknesses of Markov's ineq

How can we improve on the Markov bound?

- Do we want to bound absolute error or relative error?

- What other information from our model are we not using?

$\rightarrow E(X)$

$$P(|X - E(X)| \geq \epsilon) < \delta$$

Let's apply Markov's inequality on $|X - E(X)|$ rather than just X.

$X: E_{out}$

$E(X): E_{test}$

ϵ - bound

δ - probability

Chebyshev's Inequality

Chebyshev's Inequality: $\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \epsilon\right] \leq \dfrac{\mathrm{Var}(X)}{\epsilon^2}$ **or** $\boxed{\mathbb{P}\left[|X - \mathbb{E}[X]| \geq k\sigma\right] \leq \dfrac{1}{k^2}}$

We must have Var/stdev

Let's go back to polling, let X be the proportion of an n-sized sample that wants a proposition to pass, what are some steps that we can take here?

Chebyshev's Inequality

Chebyshev's Inequality: $\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \epsilon\right] \leq \frac{\mathrm{Var}(X)}{\epsilon^2}$ **or** $\mathbb{P}\left[|X - \mathbb{E}[X]| \geq k\sigma\right] \leq \frac{1}{k^2}$

Let's go back to polling, let X be the proportion of an n-sized sample that wants a proposition to pass, what are some steps that we can take here?

X is a sum of n individual Bernoulli distributed polls, X_i, s.t. $\mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = p.$ where p is the distribution

Recall: the variance of a Bernoulli distribution = p(1-p) < 1/4

$\sigma^2 \leq 1/4$

Since error is [0,1] bounded

$X = \sum X_i$

$\to n$

$E_{test} \pm 0.05$ w.p. 95%

Chebyshev's Inequality

Chebyshev's Inequality: $\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \epsilon\right] \leq \frac{\mathrm{Var}(X)}{\epsilon^2}$ **or** $\mathbb{P}\left[|X - \mathbb{E}[X]| \geq k\sigma\right] \leq \frac{1}{k^2}$

square term

$k > 1$

$\epsilon = 0.05$

$\delta = 0.05$

Solve for n

Let's go back to polling, let X be the proportion of an n-sized sample that wants a proposition to pass, what are some steps that we can take here?

X is a sum of n individual Bernoulli distributed polls, X_i, s.t. $\mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = p.$
where p is the distribution
Recall: the variance of a Bernoulli distribution = p(1-p) < ¼

$$\mathbb{P}[|X - p| \geq \epsilon] \leq \frac{\frac{1}{4n}}{\epsilon^2} = \frac{1}{4n\epsilon^2}$$

in terms of n

average $\frac{\text{sum}}{n}$

Hoeffding Bound

errors for points

Notice that this is an exponential bound on a sum of bounded random variables

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum(a_i - b_i)^2}\right)$$

$a_i \leq X_i \leq b_i$

$\forall i \quad a_i = 0 \quad b_i = 1$

Usually, we set the right portion to be delta (our confidence level) and then calculate the margin given the number of samples *n*.

$\delta = 0.05$

solve for ϵ

Hoeffding Bound

Notice that this is an exponential bound on a sum of bounded random variables

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum(a_i - b_i)^2}\right)$$

n

Usually, we set the right portion to be delta (our confidence level) and then calculate the margin given the number of samples *n*.

The Hoeffding bound does not directly incorporate mean or variance of the sum. Only that the individual X_i are i.i.d. random Bernoulli trials from a to b.

Hoeffding Bound

Probably - δ
Approximately - ϵ
Correct

Our form, supposing that E[Z] is our Etest and that all our trials are between 0 and 1:
(a$_i$ and b$_i$ are 0 and 1 respectively for all i)

actual error

$$Pr\left[\left|\underbrace{\frac{1}{n}\sum_{i=1}^{n} Z_i}\right. - E[Z]\right| > \epsilon\left.\vphantom{\sum}\right] \leq \delta = 2\exp\left(-2n\epsilon^2\right)$$

Error 0.01 ± 0.02
w.p 95%

$E(Z) = E_{test}$

Hoeffding Bound

Our form, supposing that E[Z] is our Etest and that all our trials are between 0 and 1:
(a_i and b_i are 0 and 1 respectively for all i)

$$Pr\Big[\Big|\frac{1}{n}\sum_{i=1}^{n} Z_i - E[Z]\Big| > \epsilon\Big] \leq \delta = 2\exp\big(-2n\epsilon^2\big)$$

Notice, that for us, confidence is "cheaper" than accuracy

$$n \geq \frac{1}{2\epsilon^2}\log\frac{2}{\delta}.$$

So we need
More datapoints
to take our bound
from Ztest to
ϵ than it
does to go
from Zt to δ

Hoeffding Bound

Our form, supposing that E[Z] is our Etest and that all our trials are between 0 and 1:

(a_i and b_i are 0 and 1 respectively for all i)

$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - E[Z]\right| > \epsilon\right] \leq \delta = 2\exp\left(-2n\epsilon^2\right)$$

Notice, that for us, confidence is "cheaper" than accuracy

$$n \geq \frac{1}{2\epsilon^2}\log\frac{2}{\delta}.$$

If we want double the confidence, we just need to increase the number of samples by some constant amount

If we want double the accuracy, $\epsilon' = \epsilon/2$, then we need 4 times the number of samples

Hoeffding Bound

95% accuracy
0.01
11598

Our form, supposing that E[Z] is our Etest and that all our trials are between 0 and 1:
(a_i and b_i are 0 and 1 respectively for all i)

$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - E[Z]\right| > \epsilon\right] \leq \delta = 2\exp\left(-2n\epsilon^2\right)$$

Usually, however, our number of samples is fixed. In general, however, you can use this when deciding how much of your data set to set aside for testing. Larger data sets don't need as big of a test set, proportionally.

$$n \geq \frac{1}{2\epsilon^2} \log \frac{2}{\delta}.$$

if you have small datasets you might need a larger testing set

Hoeffding Bound

Our form, supposing that E[Z] is our Etest and that all our trials are between 0 and 1:
(a_i and b_i are 0 and 1 respectively for all i)

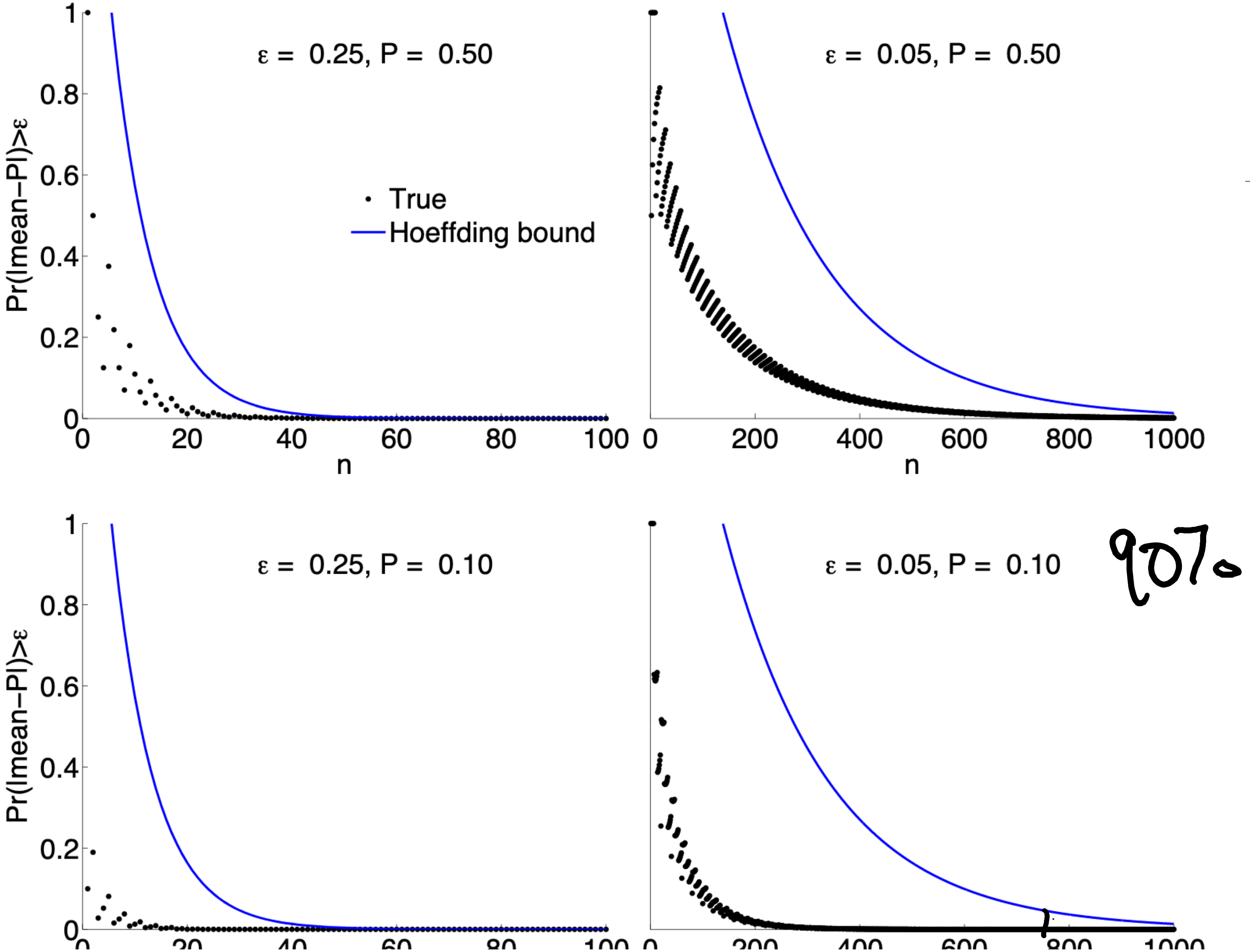
$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - E[Z]\right| > \epsilon\right] \leq \delta = 2\exp\left(-2n\epsilon^2\right)$$

Usually, however, our number of samples is fixed. In general, however, you can use this when deciding how much of your data set to set aside for testing. Larger data sets don't need as big of a test set, proportionally.

$$n \geq \frac{1}{2\epsilon^2}\log\frac{2}{\delta}. \longrightarrow \epsilon \leq \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$$

We usually want to find the bound on our error.

$E_{test} \pm \epsilon$ wp $1-\delta$



Hoeffding Bound

In general, the actual performance is better than the Hoeffding bound

relatively tight bound.

Better w/ higher n.

Hoeffding Bound

While we may choose our modeling approach, the ML algorithm itself is trialing multiple possible models.

For example, since neural networks are heavily impacted by their random start position, we may want to make several models on the test data and select which one is worse.

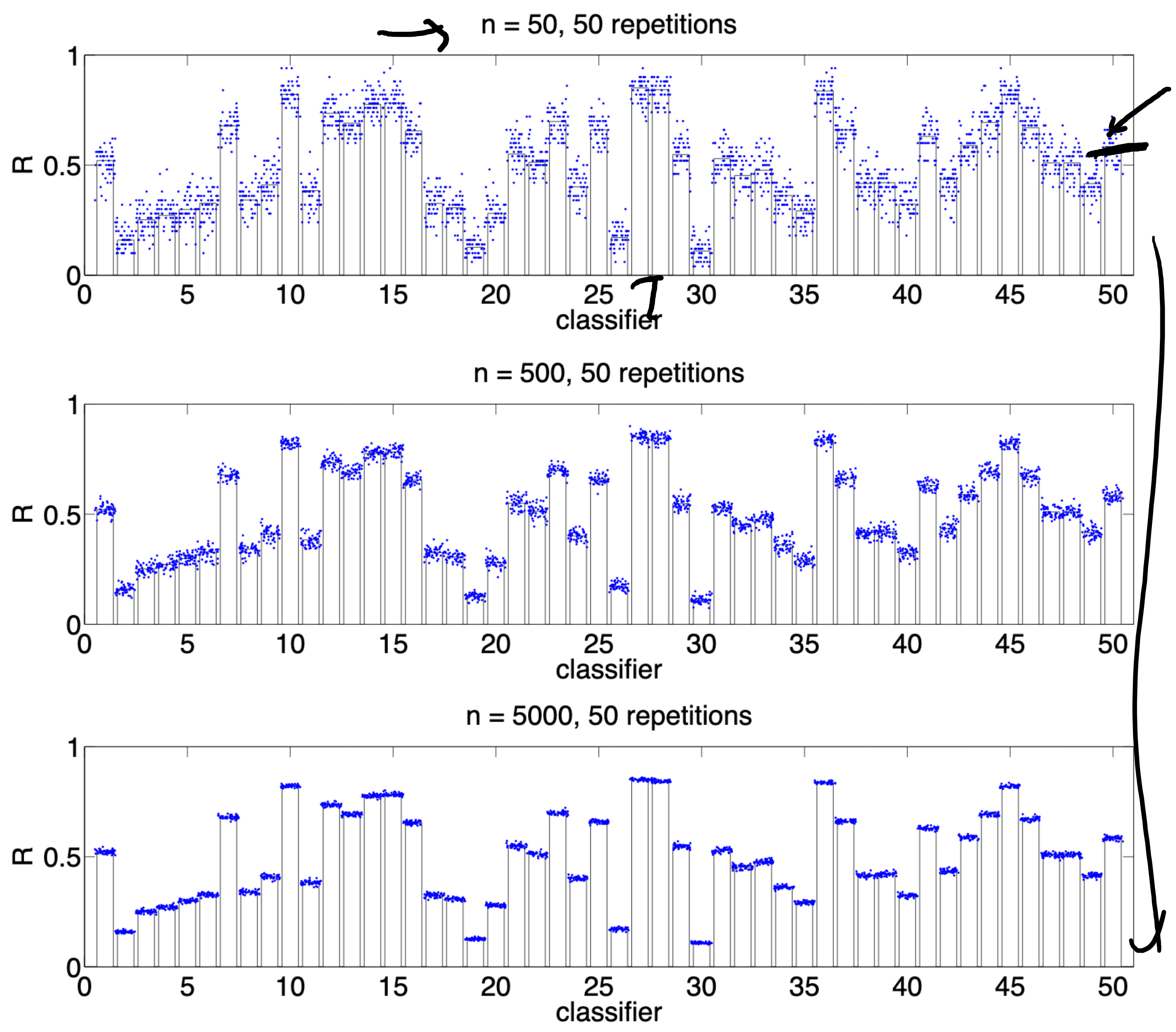
We want to reduce the number of times we run on the test set, but it may not be possible to reduce this number to one.

We then want to think: if we want an overall error of δ across G hypotheses, then each hypothesis needs to have a certainty level of: $\delta/\mathcal{G}$

This makes our new error: $\epsilon = \sqrt{\frac{1}{2n}\log\frac{2|\mathcal{G}|}{\delta}}$

if G is large

$\log(G+1) \simeq \log(G)$



Experimental Evidence

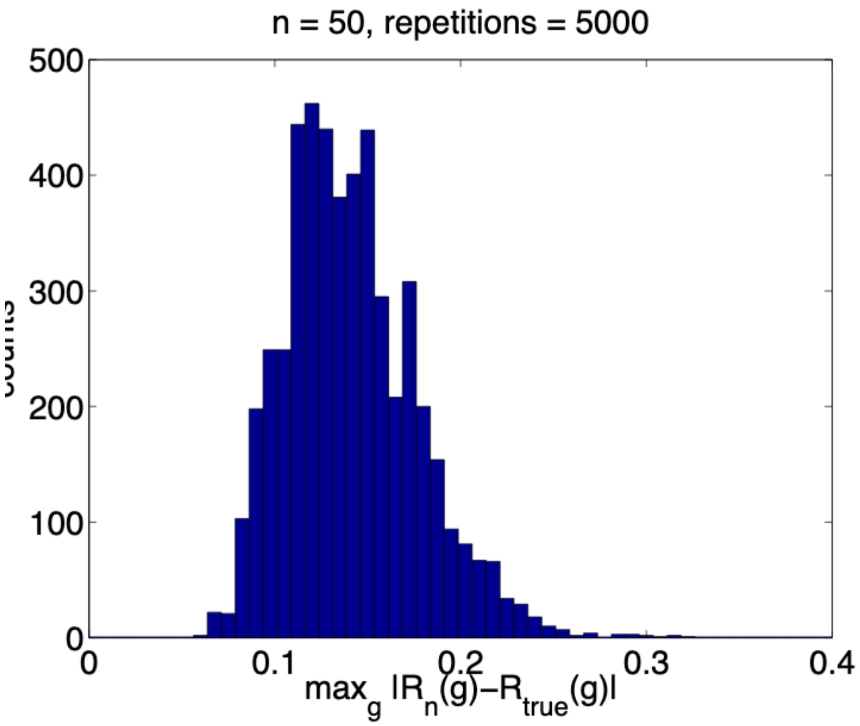
Here we've selected 50 random linear models that try and estimate linear data.

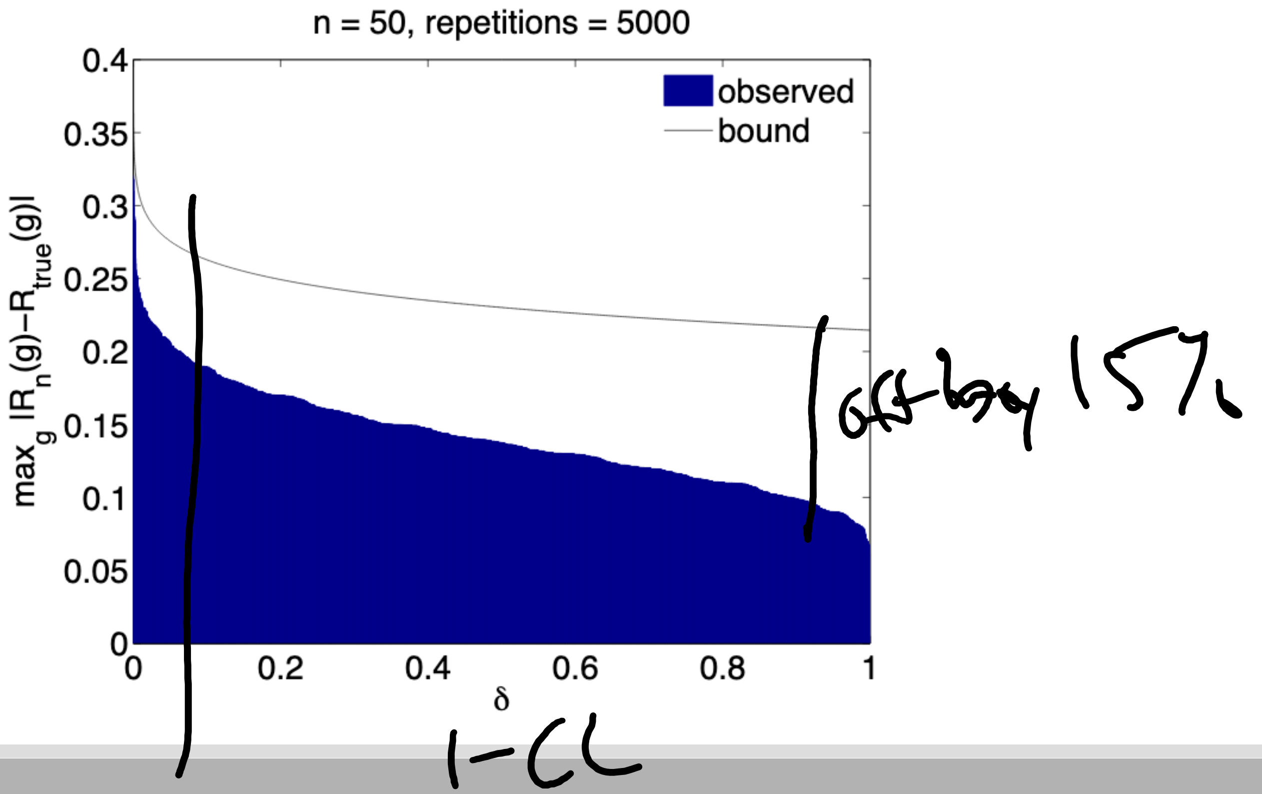
With greater n, we can see that our estimate for the error gets closer to the true error

Experimental Evidence

Here, we see the average differences between predicted an actual

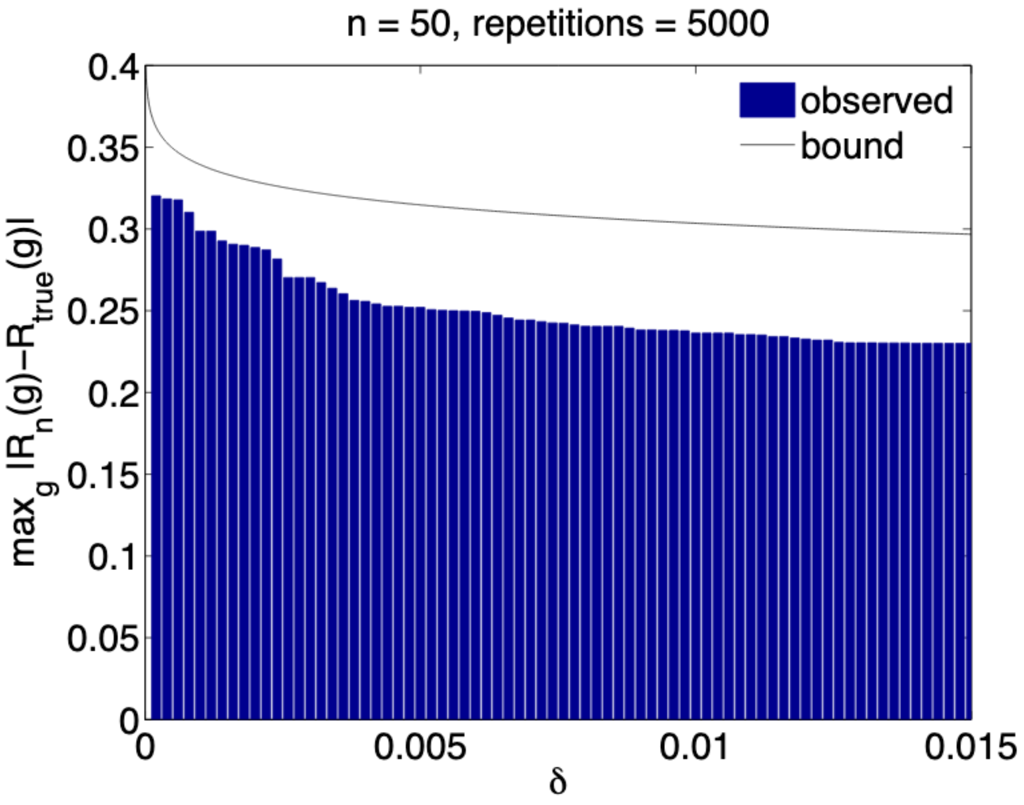
Our Hoeffding bound value doesn't seem to be very good





Experimental Evidence

However, when our confidence is high, we see that the bound on our expected error becomes a closer fit to the data itself



97-99.9+%

Structural Risk Minimization

$G_1 = kNN$

$G_2 = LR$

The Hoeffding bound is a way of mathematizing the bias-variance tradeoff

Build multiple sets of implementations G, each getting more complicated than the last and such that: $G_i \subseteq G_{i+1}$

For example, in your first iteration, consider only linear models, find the best model and calculate its expected error subject to:

$$2\sqrt{\frac{1}{2n}\log\frac{2|\mathcal{G}|}{\delta}}$$

small for early simple model

Then, keep growing G until it encapsulates all possible models

This is a cross-method approach to cross-validation – **not testing!**

Choose the hypothesis gi which has the lowest $E_{cv} + 2\sqrt{\frac{1}{2n}\log\frac{2|\mathcal{G}|}{\delta}}$

VC Dimension – variable

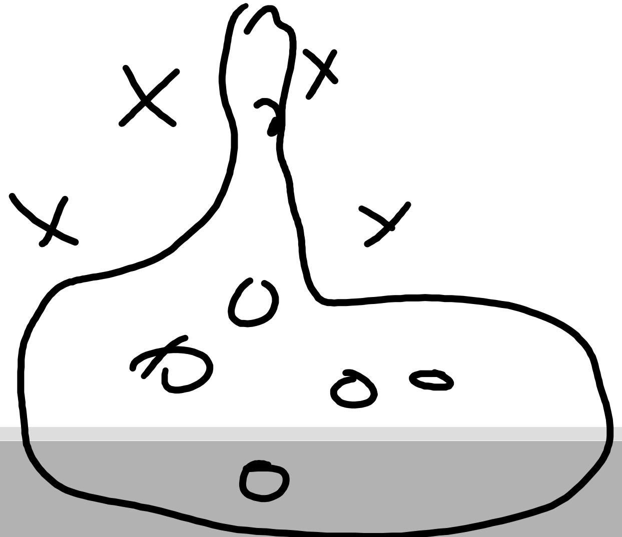
How do we measure the "complexity" of a model?

We can use the VC dimension, which represents the smallest data set which a model can fit with zero training error. no matter the points

Similar to the degrees of freedom for a linear model

degrees of freedom

What is the VC dimension for a SVM with
polynomial kernel of degree p and D features? $\binom{D+p-1}{p}$ ✓



VC Dimension

How do we measure the “complexity” of a model?

We can use the VC dimension, which represents the smallest data set which a model can fit with zero training error.

Similar to the degrees of freedom for a linear model

What is the VC dimension for a SVM with
polynomial kernel of degree p and D features? $\binom{D+p-1}{p}$

We can then apply risk minimization with a new error bound

$$\mathrm{E}_{\mathrm{cv}} + \sqrt{\underbrace{\frac{VC[\mathcal{G}]}{n}\left(\log \frac{n}{VC[\mathcal{G}]} + \log 2e\right)}_{\text{complexity penalty}} + \underbrace{\frac{1}{n}\log\frac{4}{\delta}}_{\approx\ \text{Hoeffding bound}}}.$$ with probability δ

Types of errors

$$acc = \frac{TP+TN}{n}$$

Prediction

Ground Truth	cancer ~~Edge~~	no cancer ~~Not edge~~
cancer ~~Edge~~	True Positive	False Negative
no cancer ~~Not Edge~~	False Positive	True Negative

Cancer

very costly

cancer

Two parts to each: whether you got it correct or not, and what you guessed. For example for a particular pixel, our guess might be labelled…

True Positive

Did we get it correct?
True, we did get it correct.

What did we say?
We said 'positive', i.e. edge.

TP
TN
FP
FN

or maybe it was labelled as one of the others, maybe…

False Negative

Did we get it correct?
False, we did not get it correct.

What did we say?
We said 'negative, i.e. not edge.

Sensitivity and Specificity

Count up the total number of each label (TP, FP, TN, FN) over a large dataset. In ROC analysis, we use two statistics:

Minimize? what

Sensitivity = $\dfrac{TP}{TP+FN}$

Can be thought of as the likelihood of spotting a positive case when presented with one.

Or… the proportion of edges we find.

if we go how can do we catch it

Specificity = $\dfrac{TN}{TN+FP}$

Can be thought of as the likelihood of spotting a negative case when presented with one.

Or… the proportion of non-edges that we find

if we don't do we let it go

$$\text{Sensitivity} = \frac{TP}{TP+FN} = ?\ 1/3$$

$$\text{Specificity} = \frac{TN}{TN+FP} = ?\ 1/5$$

Prediction

Sensitivity

		1	**0**
Ground Truth	**1**	60 90	30 0
	0	80 100	20 0

60+30 = 90 cases in the dataset were class 1

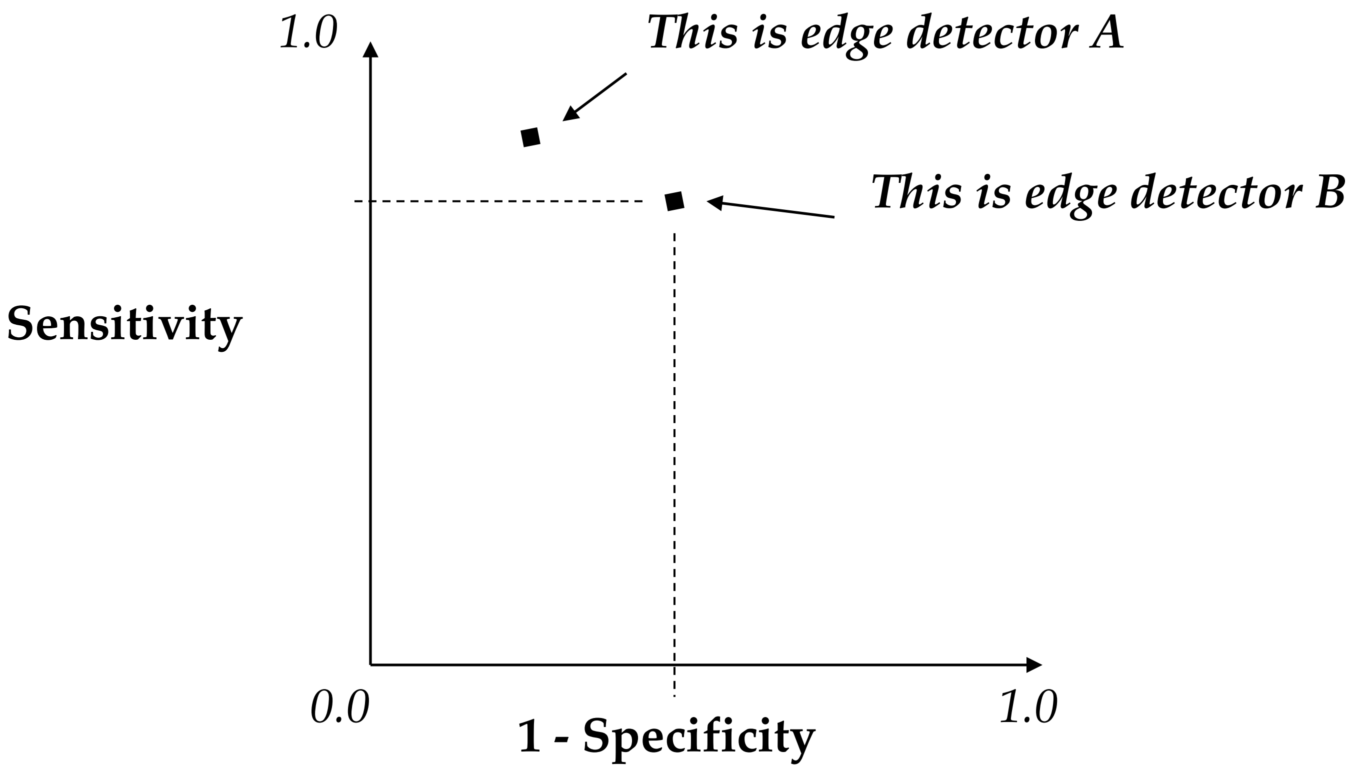
80+20 = 100 cases in the dataset were class 0

Sensitivity = 1 00

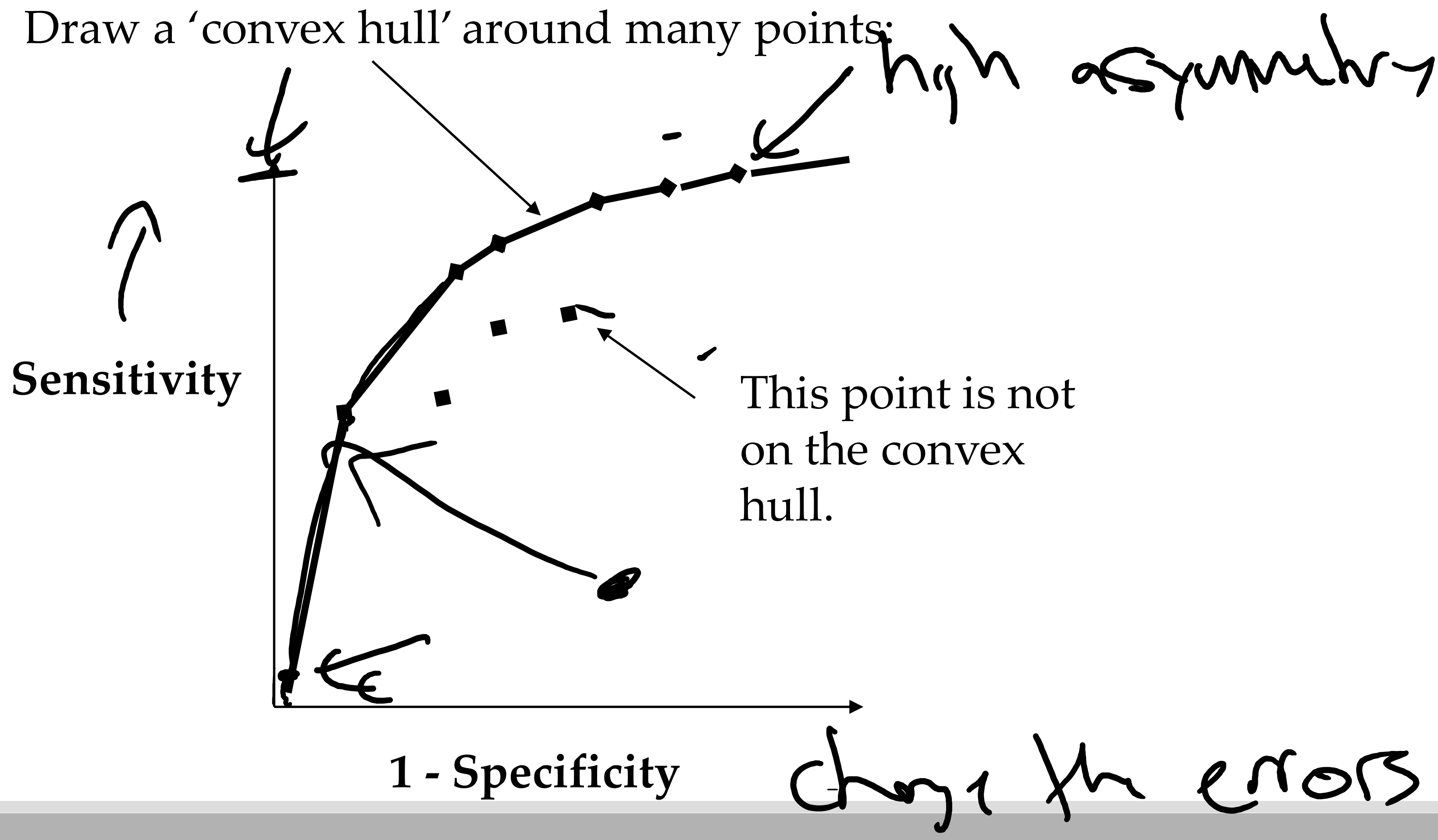
specificity = 0

The ROC space

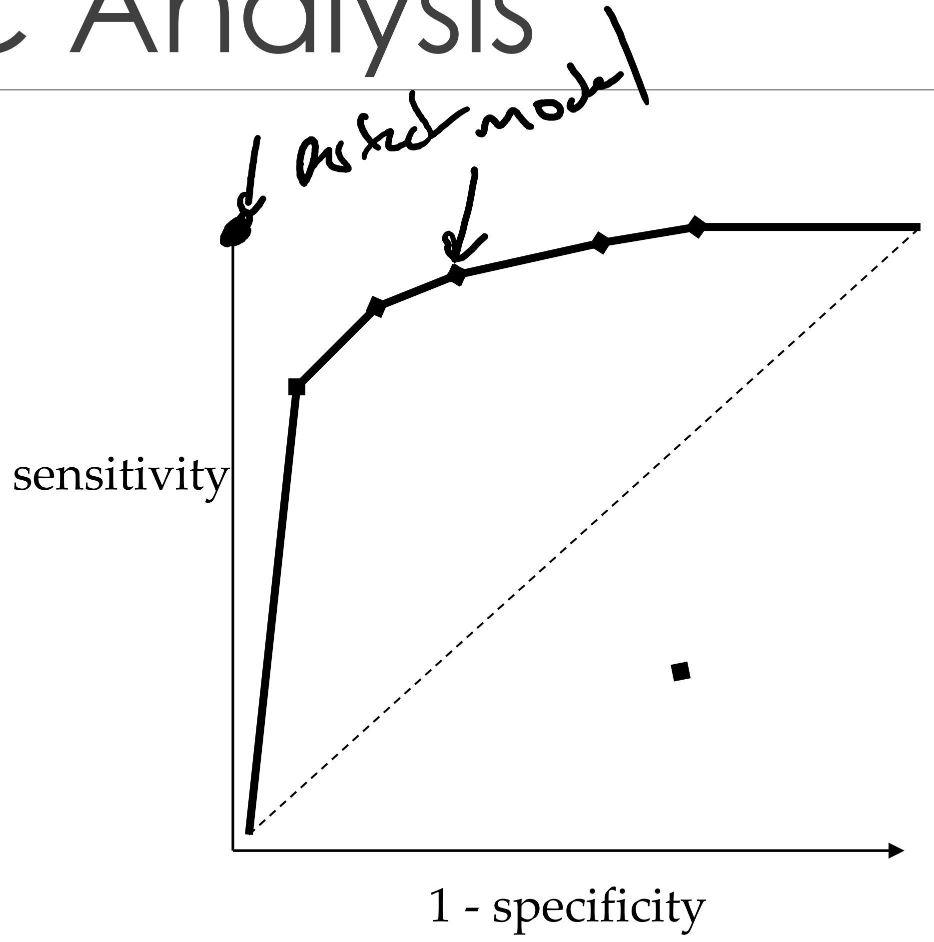
(Receiver Operating Characteristic)



The ROC Curve



ROC Analysis



All the optimal classifiers lie on the convex hull.

Which of these is best depends on the ratio of edges to non-edges, and the different cost of misclassification