Announcements

Regular lecture on Wed this week

Will finish grading in a couple of days

Group project "model" specs due by the 13th!

Easy and Hard Predictions

Why Accuracy and Error Are Misleading

CHIDATA

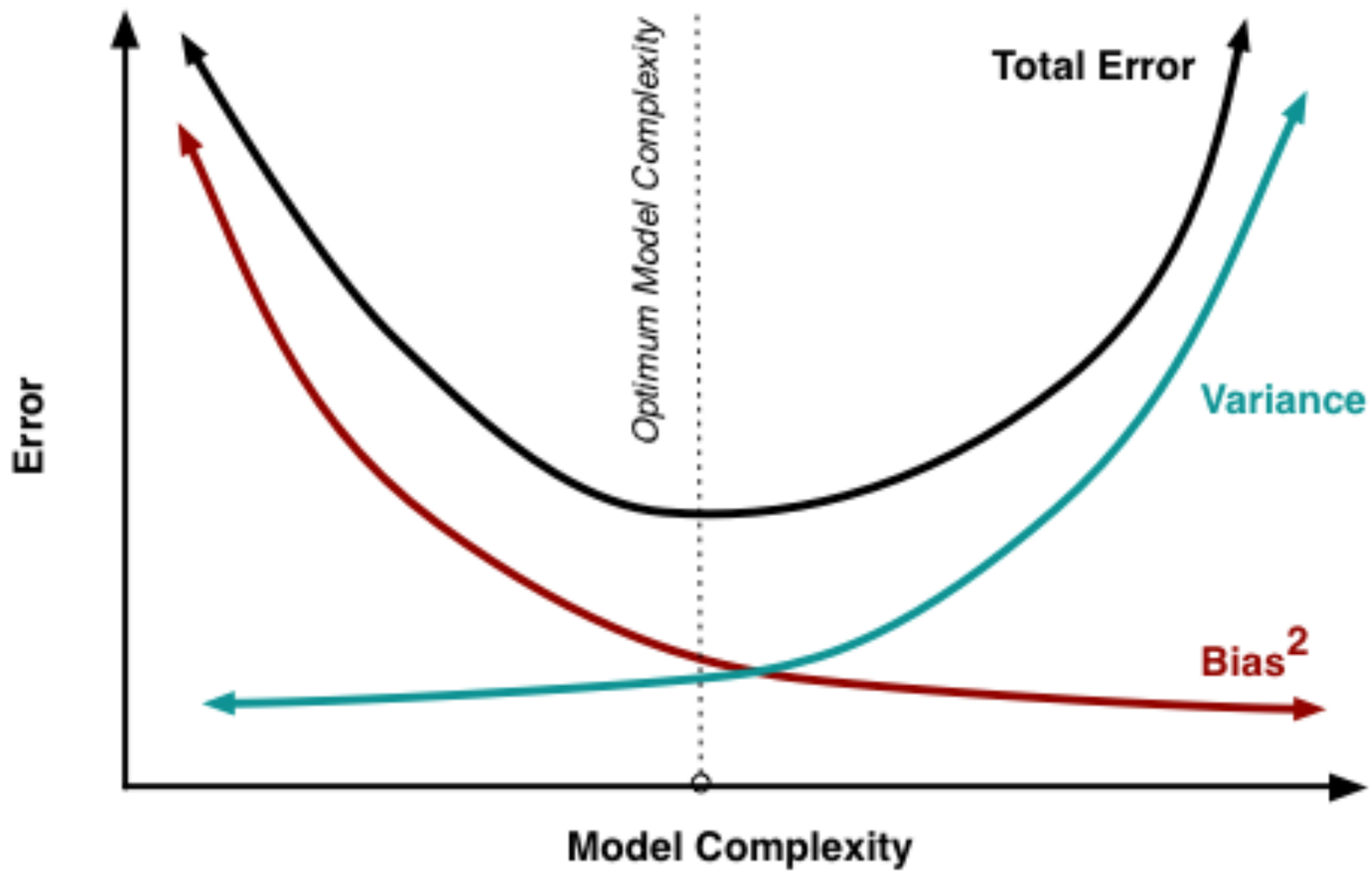
Recap: Simulation v.s. Model Fitting

Fitting: Observe xi, yi, **infer theta**

Regression Problems -> Continuous Y

Classification Problems -> Discrete Y

Recap: Bias-Variance Tradeoff



Fraud Detection

Task: Build a model to determine whether a transaction at a bank is fraudulent or not.

Classification Problems -> Discrete Y

Suppose: 99.97% of transactions are honest and 0.03% are fraudulent.

How good is random(ish) guessing?

Look at some data-independent baselines

Suppose: 99.97% of transactions are honest and 0.03% are fraudulent.

Classifier: With 50% probability pick Fraud, 50% Not Fraud

(Bad) Idea 1. Even Guess

Suppose: 99.97% of transactions are honest and 0.03% are fraudulent.

Classifier: With 50% probability pick Fraud, 50% Not Fraud

What is the accuracy?

Accuracy = 50% = 0.5*.9997 + 0.5*0.0003

Predict Not Fraud Correctly **Predict Fraud Correctly**

(Bad) Idea 2. Best Guess

Classifier: With **p** probability pick Class 1, **(1-p)** Class 2

Accuracy = p* (class1 %) + (1-p)* (class 2 %)

Accuracy = class 2 % + p* (class1 % - class2 %)

Accuracy = c2 + p* (c1-c2)

Maximized at: p = 1 if c1 > c2
p = 0 if c2 > c1

How good is random(ish) guessing?

Suppose: 99.97% of transactions are honest and 0.03% are fraudulent.

Classifier: Pick highest likelihood class

99.97% Accuracy

Why is this a misleading statement?

How good is random(ish) guessing?

Suppose: 99.97% of transactions are honest and 0.03% are fraudulent.

Classifier: Pick highest likelihood class

0% Accurate

True=
Fraud

100% Accurate

True =
Not Fraud

False Positives v.s False Negatives

False Positives: Predict fraud when not fraud

False Negatives: Predict not fraud when actually fraud

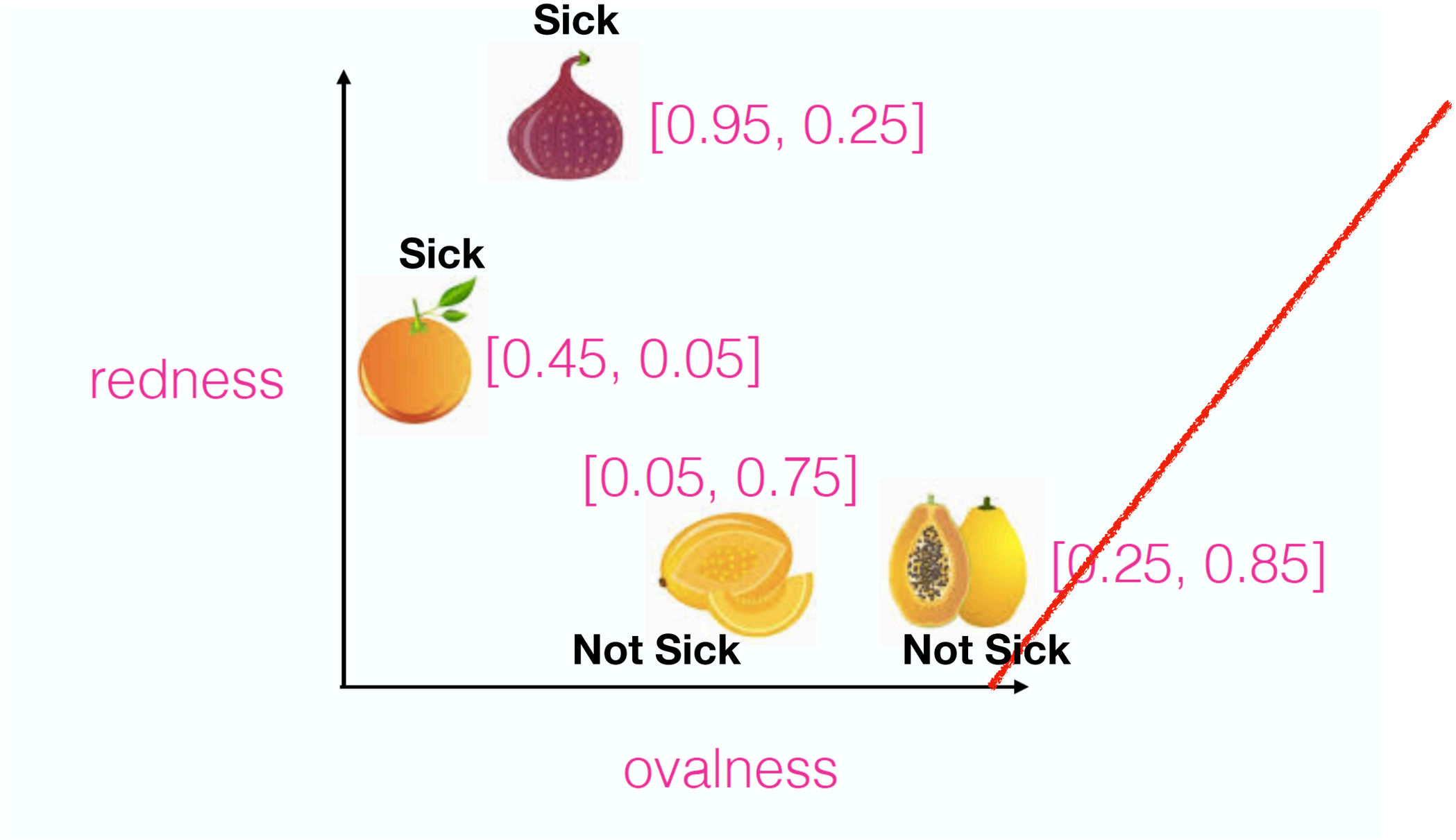
Can tradeoff false positives for false negatives!

Logistic Regression

	X [redness, how oval]	Prediction
	[0.95, 0.25]	0.96 > a ~~ 1
	[0.25, 0.85]	0.1 > a ~~ 0
	[0.45, 0.05]	0.87 > a ~~ 1
	[0.05, 0.75]	0.14 > a~~ 1

Set threshold a for desired tradeoff point

All Classifiers Have This Property



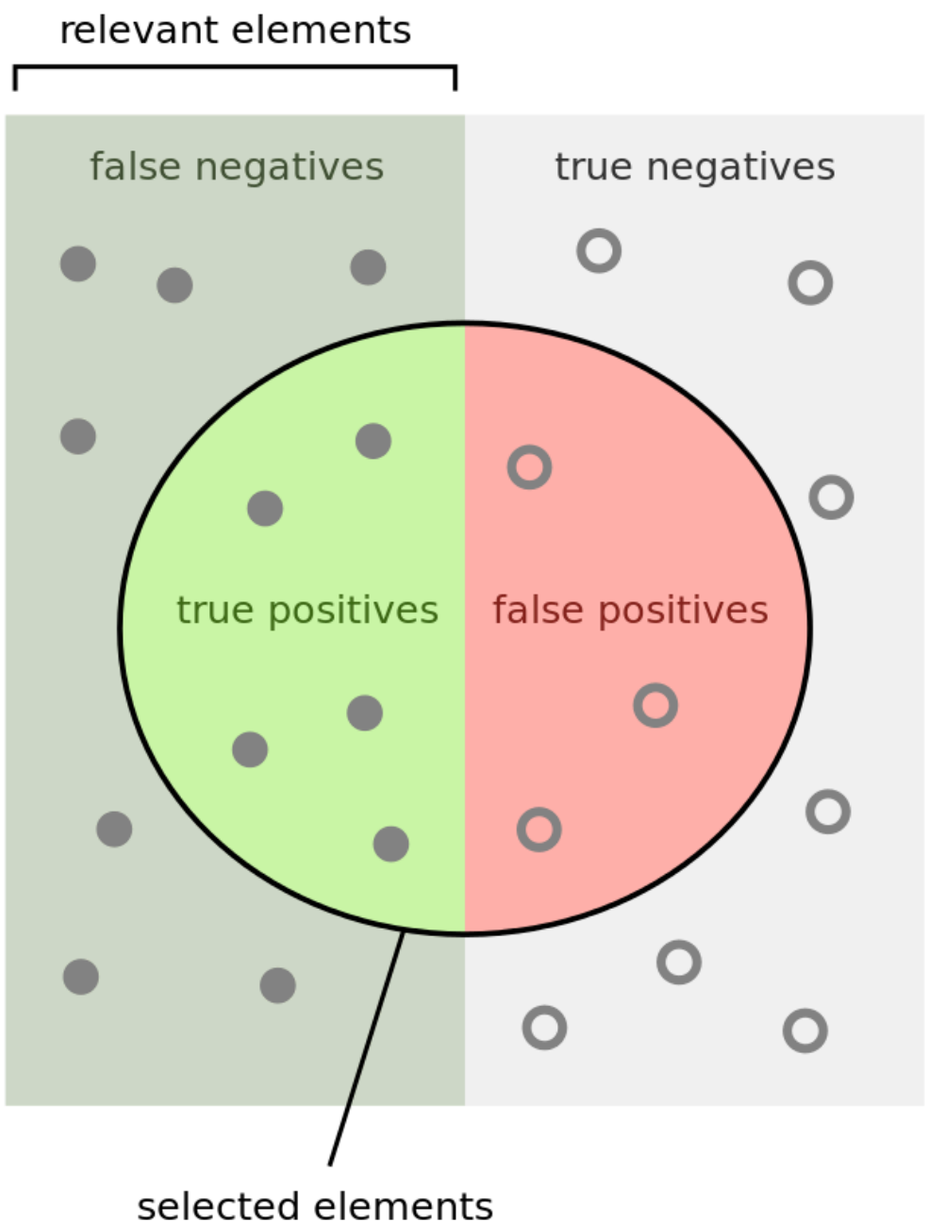
Precision and Recall

Precision: Fraction of truly detected among all detected.

tp/(tp+fp)

Recall: Fraction of the total amount of detected that were actually detected.

tp/(tp+fn)



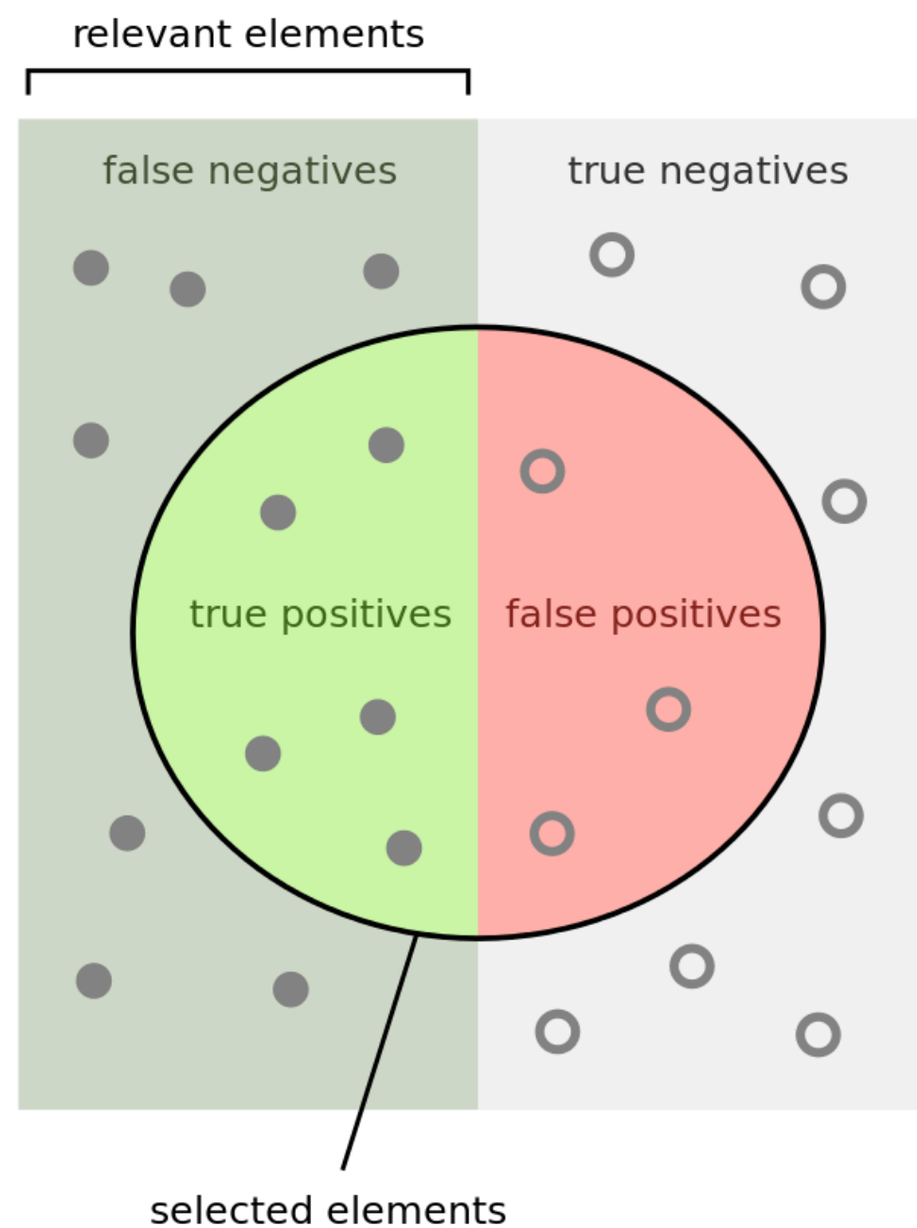
Precision and Recall

Precision: Fraction of truly detected among all detected.

$$\frac{\#detect\_fraud}{(\#detect\_fraud + \#wrong)}$$

Recall: Fraction of the total amount of detected that were actually detected.

$$\frac{\#detect\_fraud}{(\#detect\_fraud + \#missed)}$$



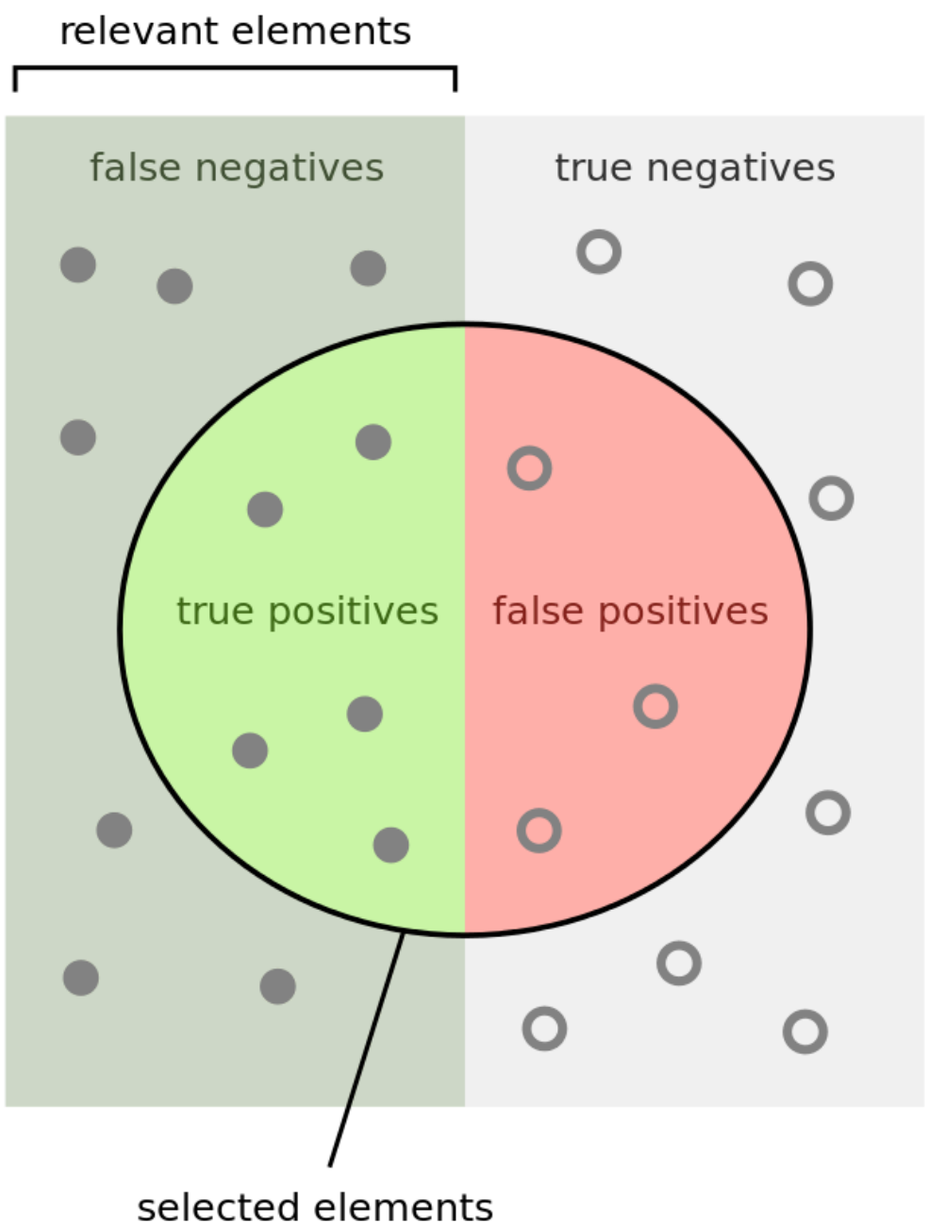
Precision and Recall

Precision: Fraction of truly detected among all detected.

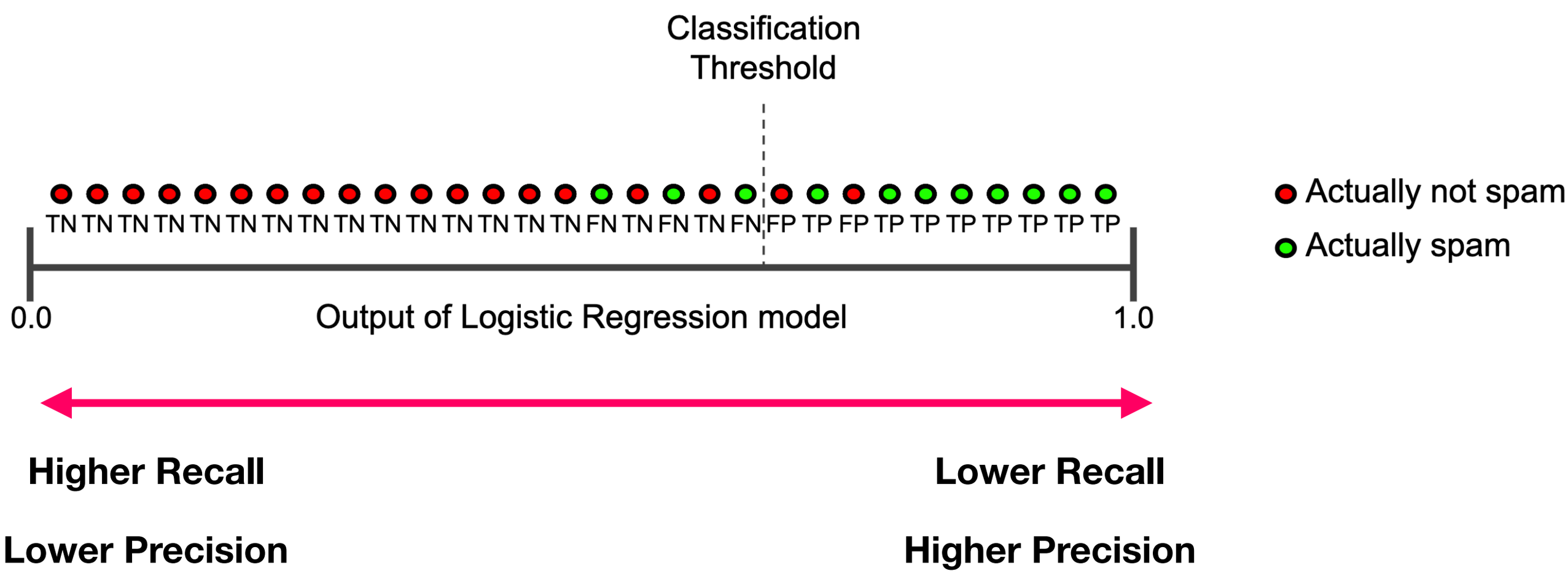
False positive rate

Recall: Fraction of the total amount of detected that were actually detected.

False negative rate



Visualization



Can flip them too!

False Positives: Predict **not fraud** when not fraud

False Negatives: Predict fraud when actually fraud

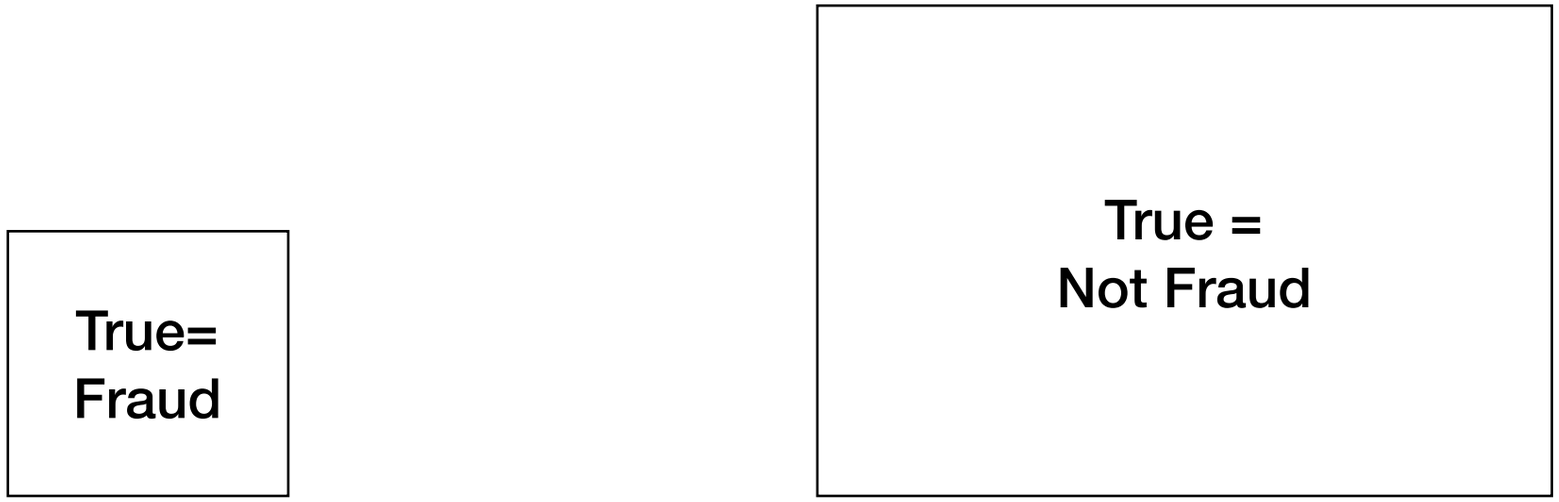
Can tradeoff false positives for false negatives!

Dealing With Imbalances

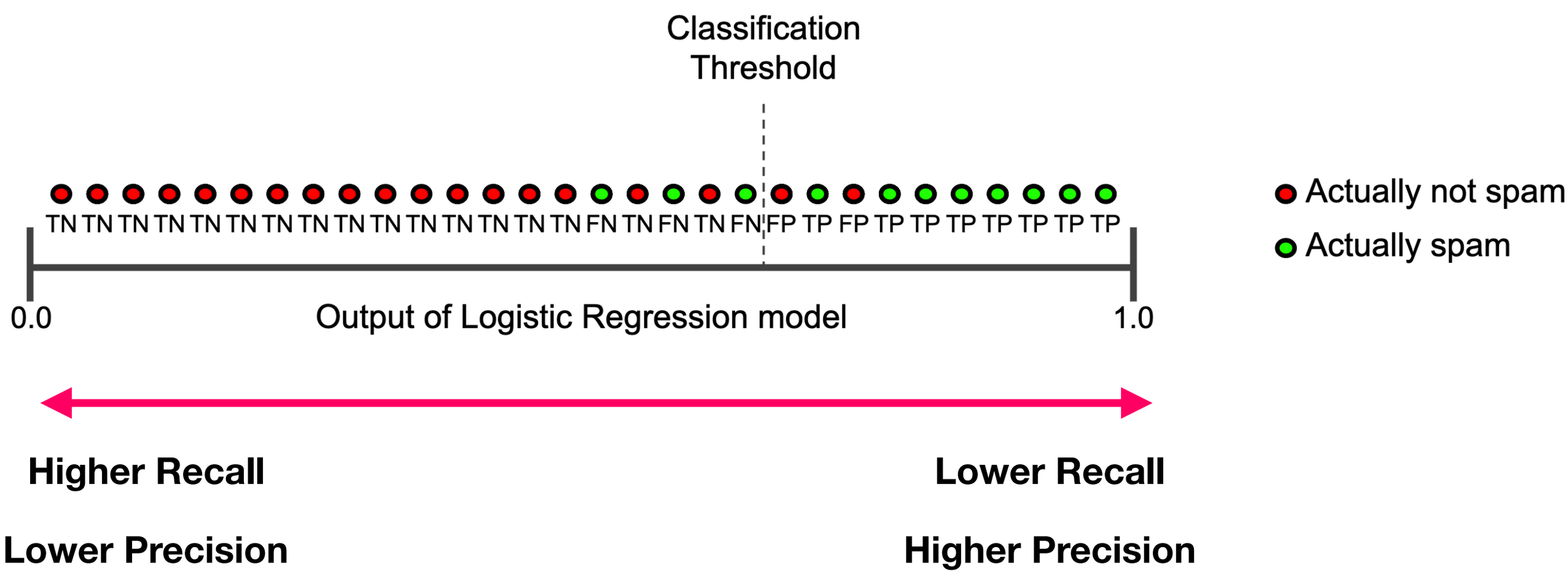
Suppose: 99.97% of transactions are honest and 0.03% are fraudulent.

Measure: Per-Class Precision and Recall

A good classifier is roughly balanced in P/R per class

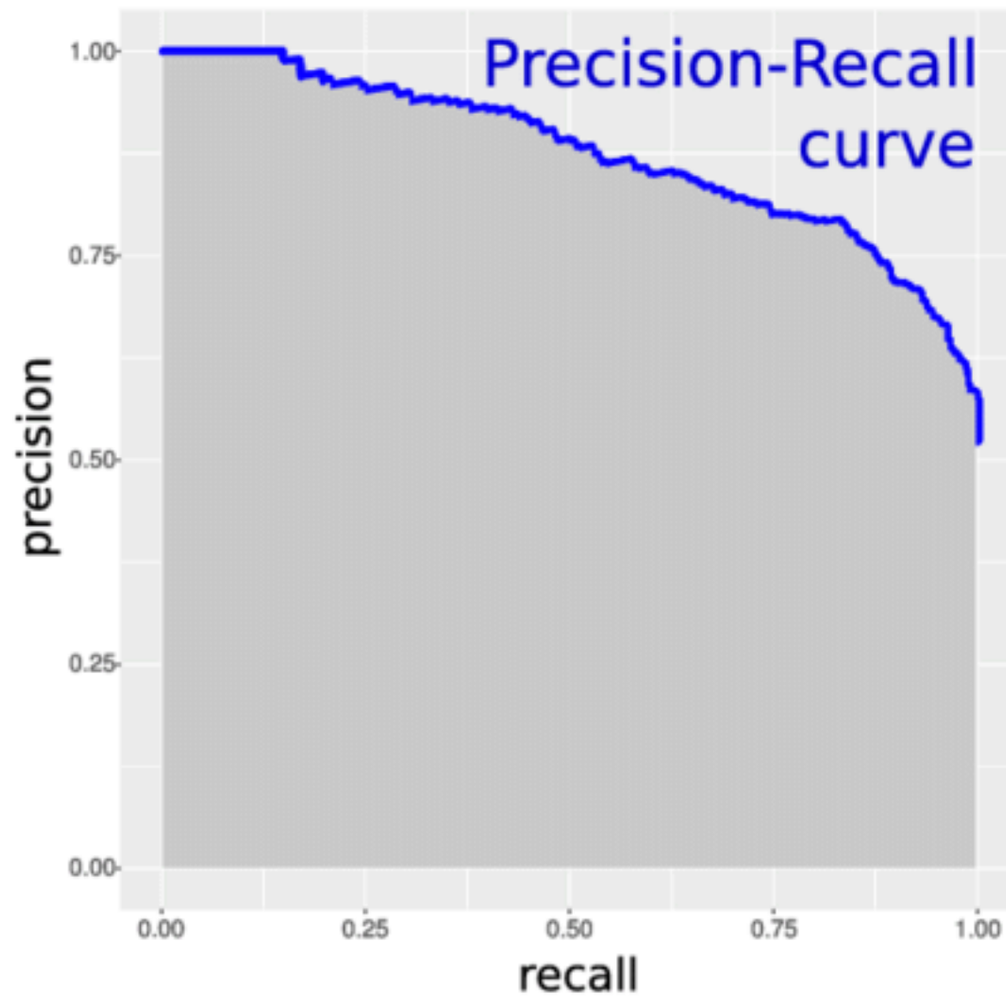


Dependency on the Threshold?



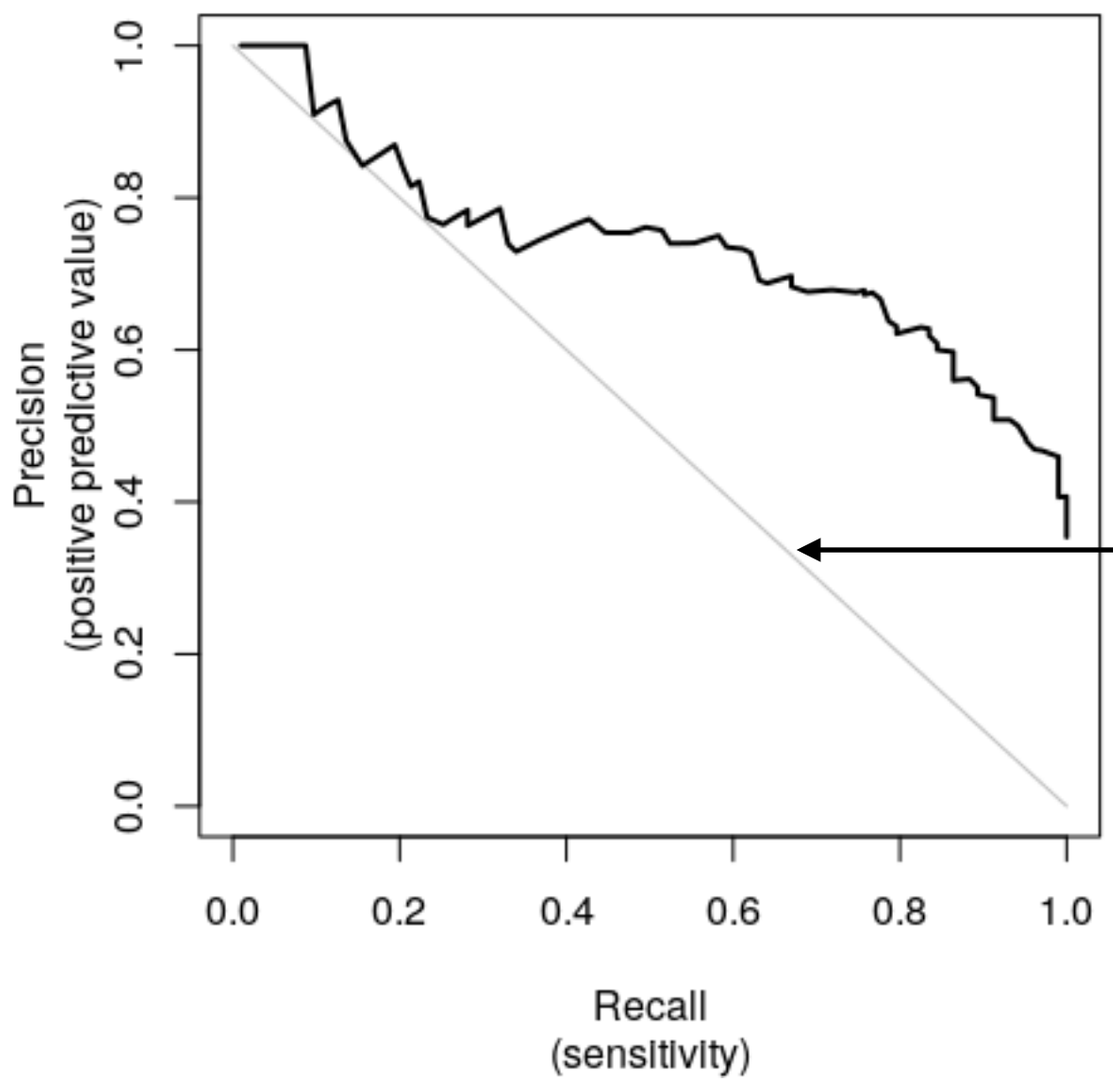
Precision Recall Curves

For all possible classification thresholds plot the P/R Curve



Precision Recall Curves

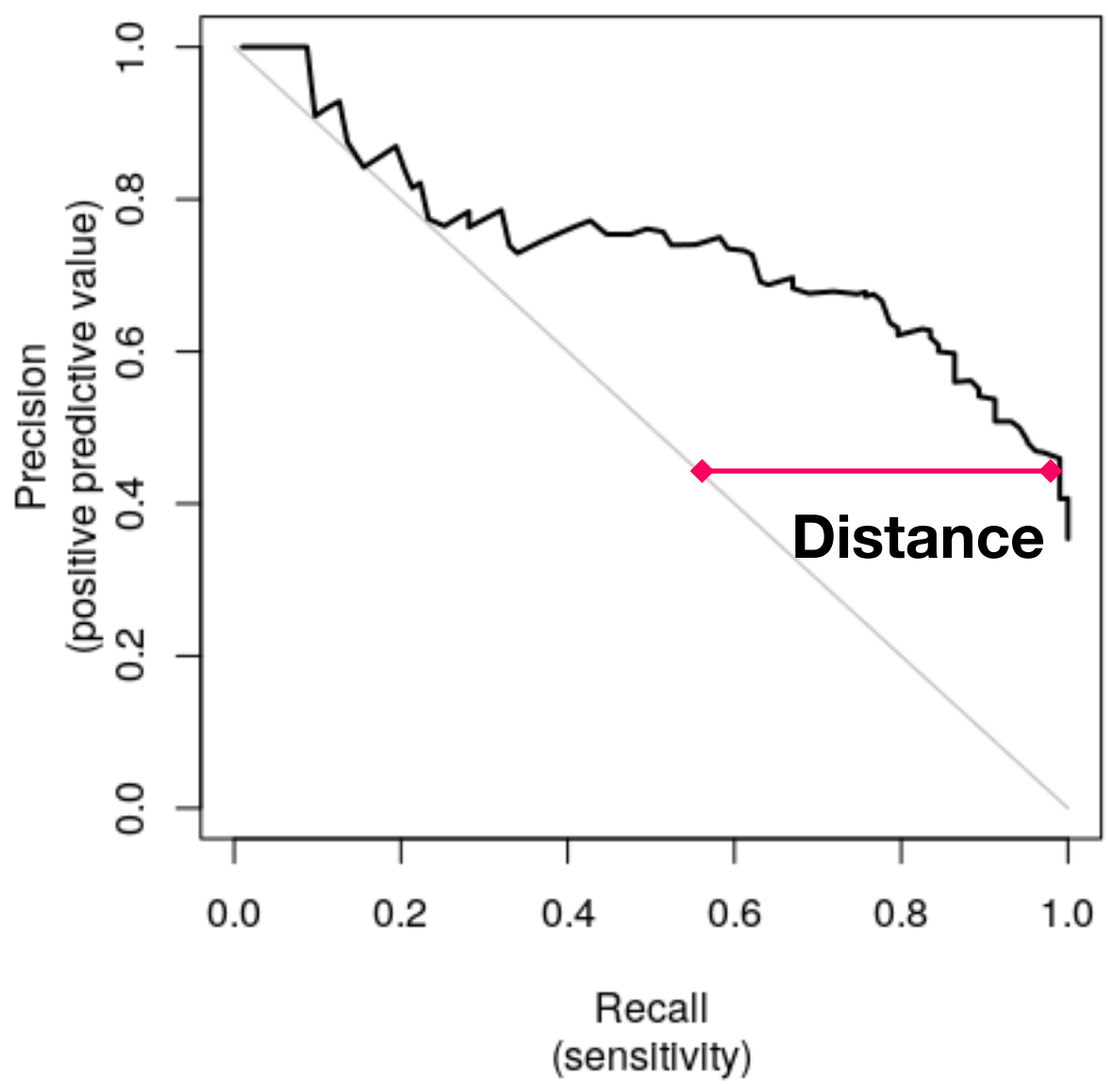
For all possible classification thresholds plot the P/R Curve



Random guessing!!!

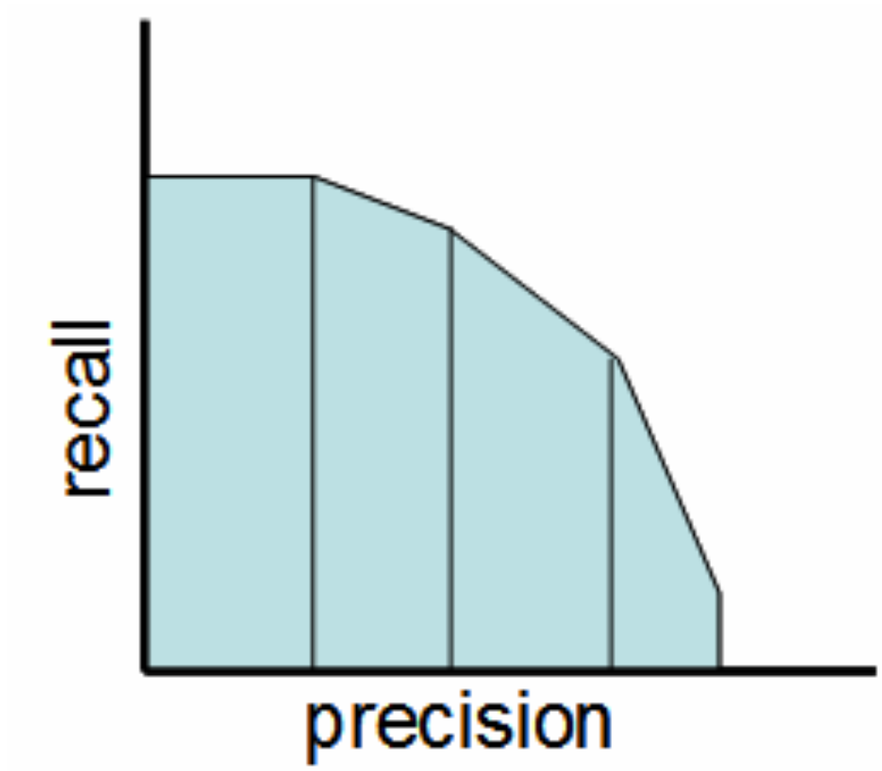
Precision Recall Curves

For all possible classification thresholds plot the P/R Curve



Precision Recall Curves

Area-Under-Curve (AUC)

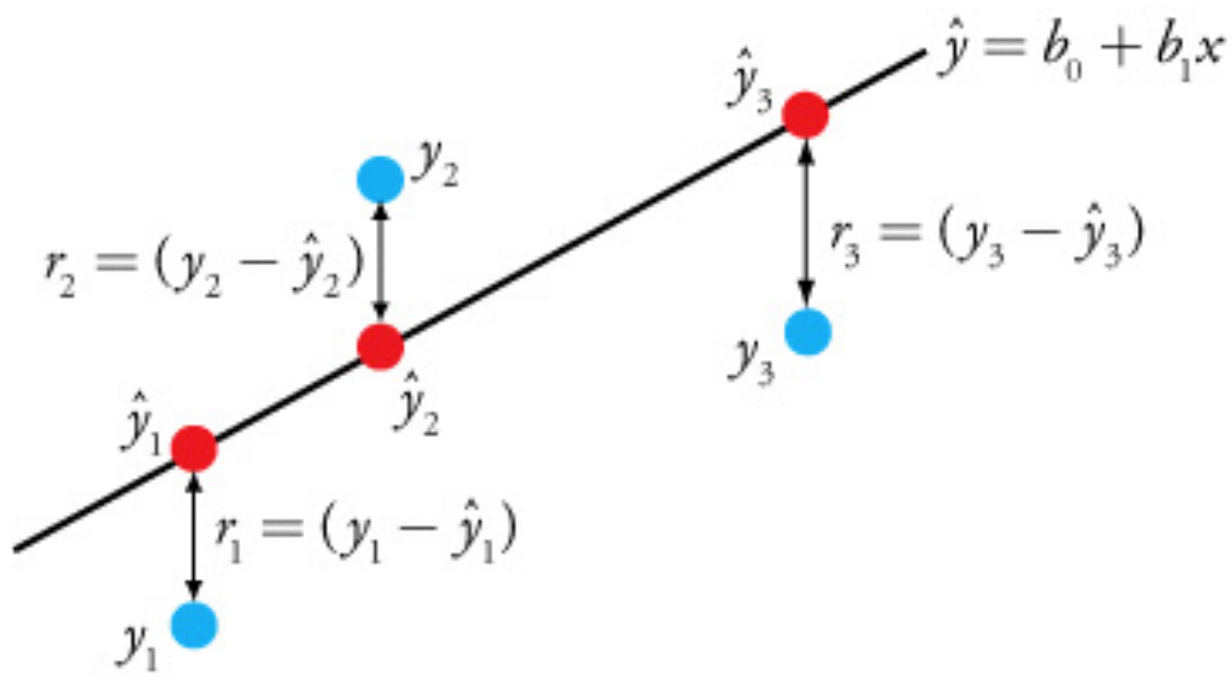


0.5 = Random Guessing

> 0.5 Better than random

Regression Problems

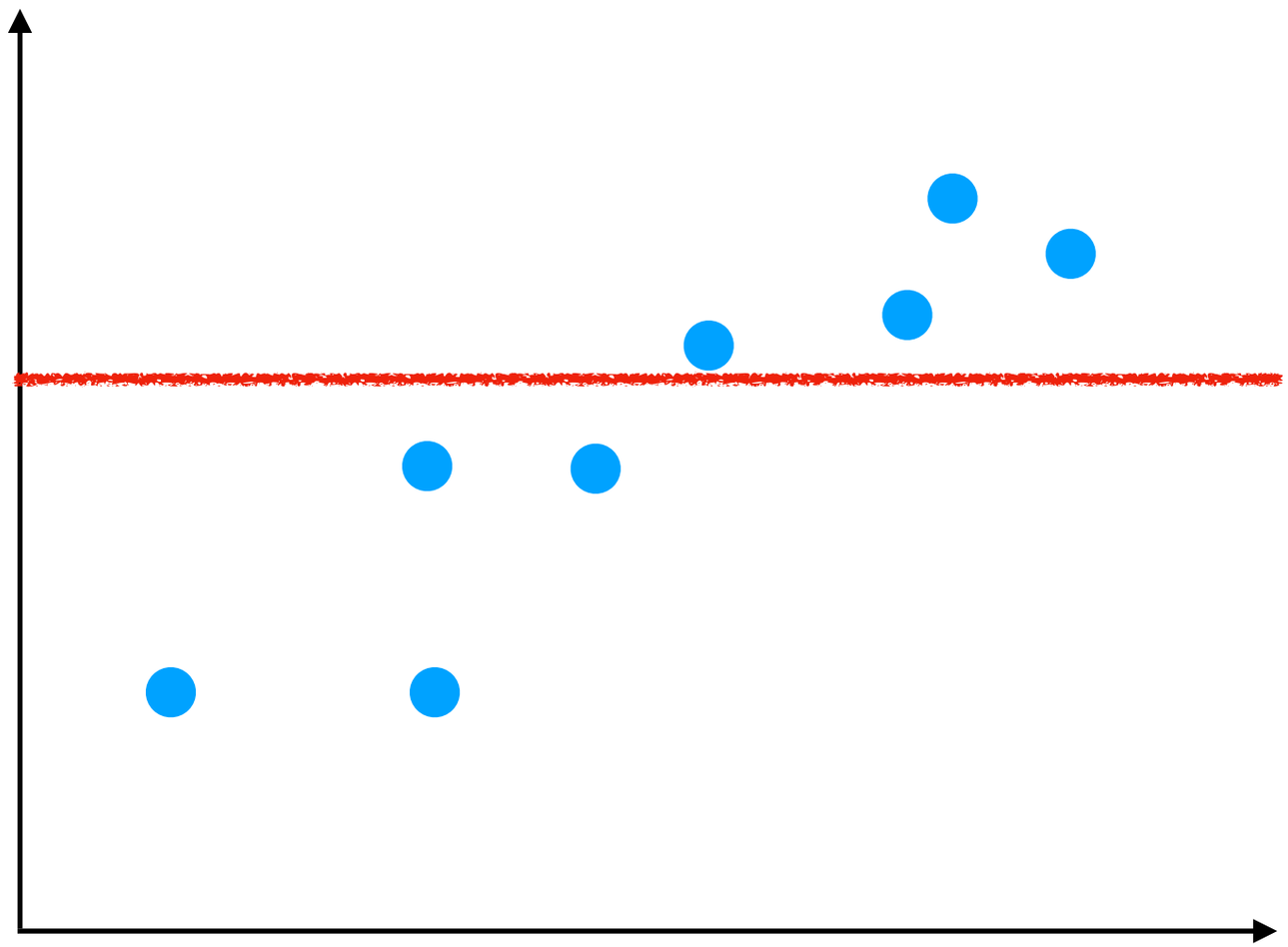
Measure errors in regression problems



1-parameter regression

Data-independent baselines

$$\min_{c} \sum_{i=1}^{N} (y_i - c)^2$$



1-parameter regression

$$f(c) = \sum_{i=1}^{N} (y_i - c)^2$$

$$f'(c) = \sum_{i=1}^{N} -2(y_i - c)$$

$$0 = \sum_{i=1}^{N} -2(y_i - c)$$

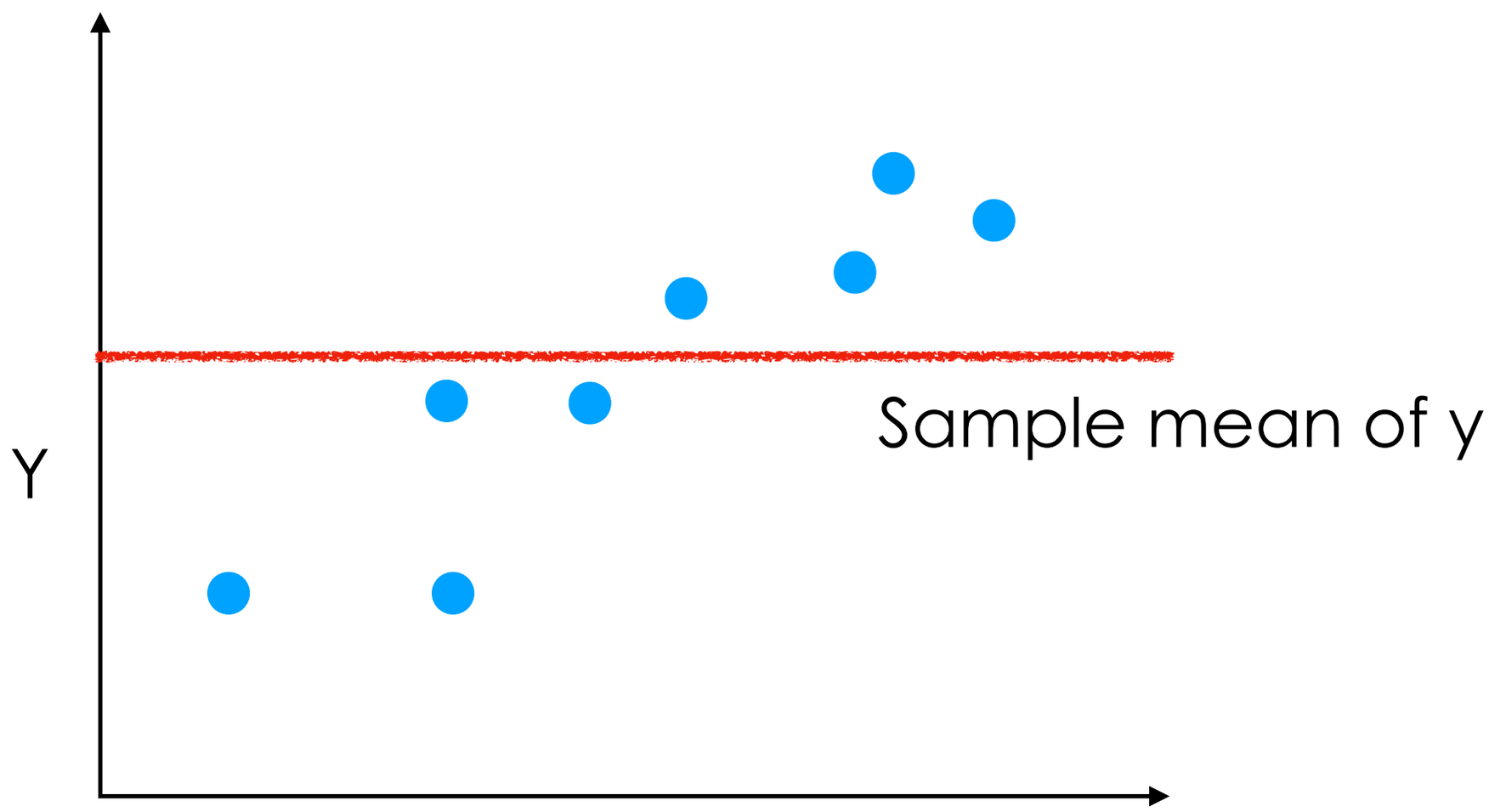
$$0 = -2\sum_{i=1}^{N} y_i + 2Nc$$

$$\boxed{c = \frac{1}{N}\sum_{i=1}^{N} y_i}$$

C is the sample mean!!!

1-parameter regression

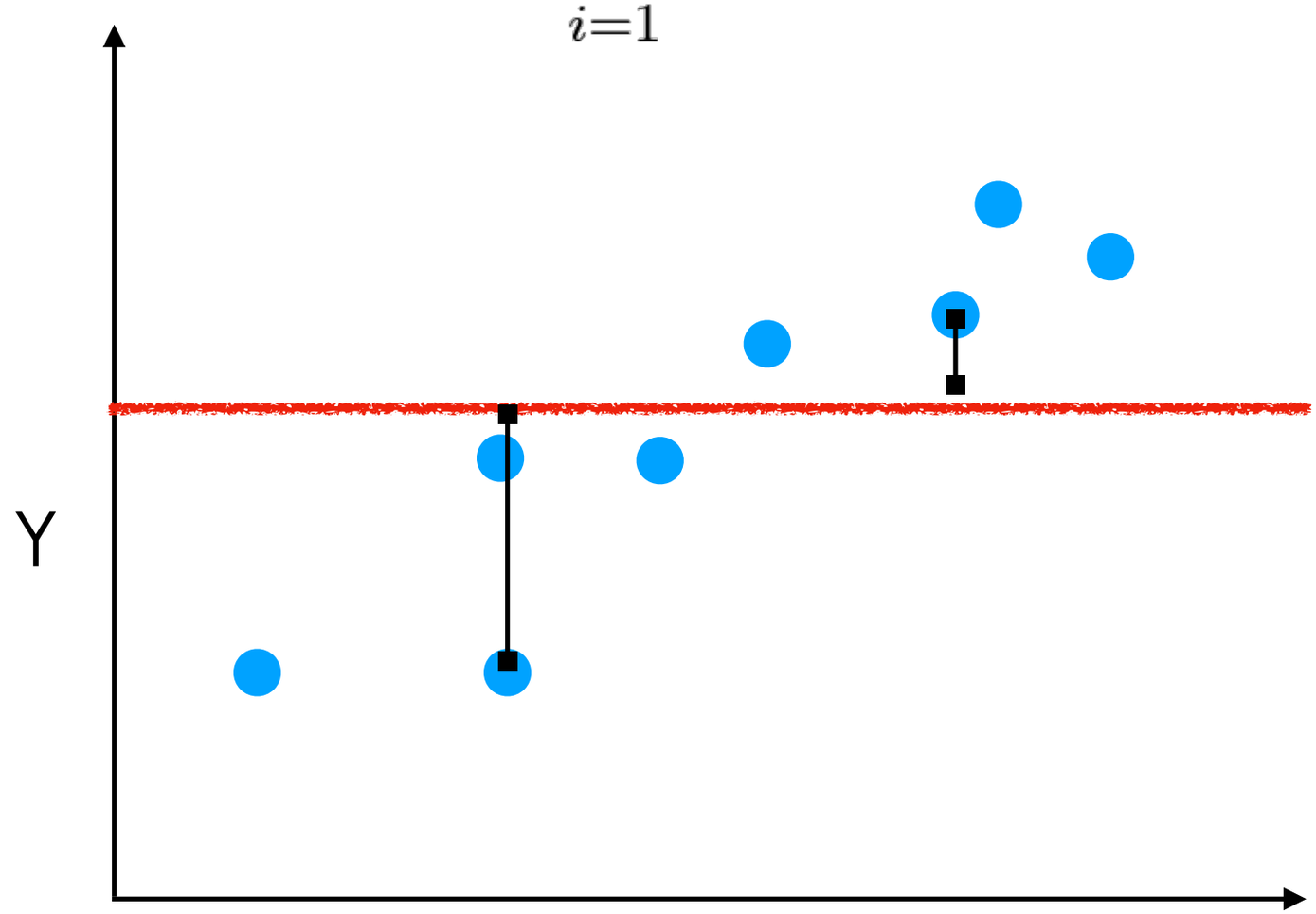
Best constant prediction is the mean of y!



1-parameter regression

What about the "error" in estimation?

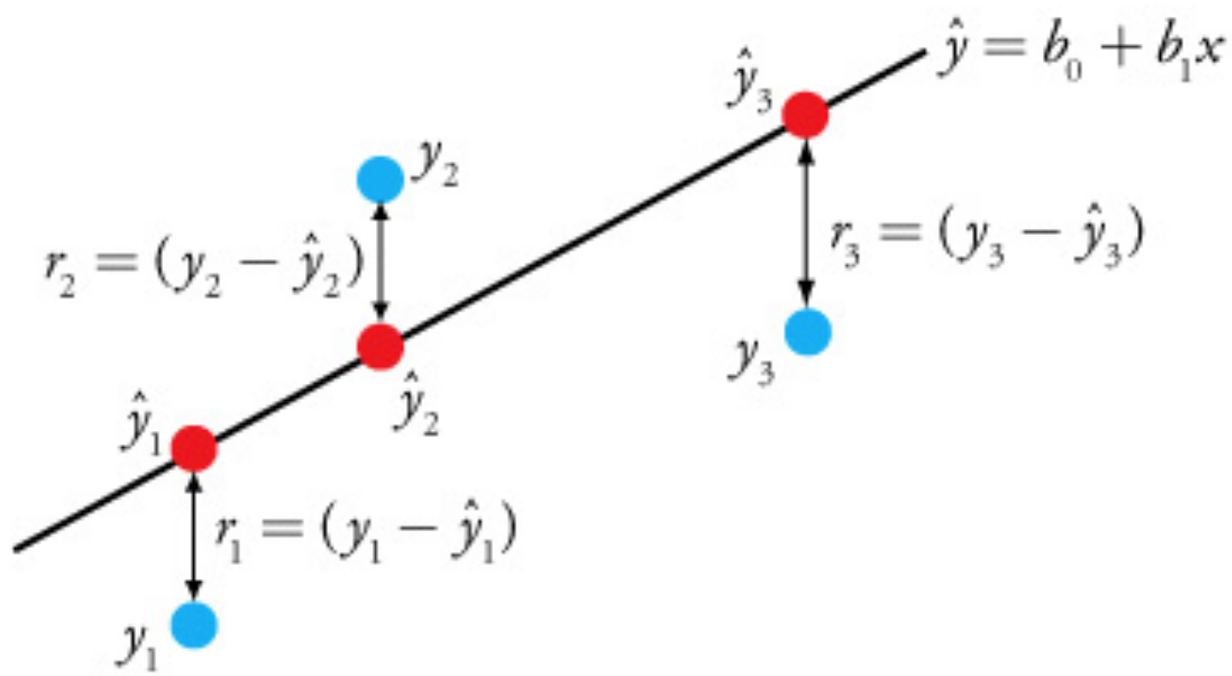
$$\sum_{i=1}^{N}(y_i - \mu_y)^2$$



Proportional to the variance of y!

Benchmarking Regression

Compare regression results to a mean guess



R2 Score

Calculate the actual error compared to the error in a mean guess

Error

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Variation

Score = 1 (perfect prediction)

Score = 0 (no better than the mean)

Score < 0 (worse than the mean)

Explained Variation

$$1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

“Explained” variation

How much of the variation in the data is explained by the predictor.

Explained Variation

$$1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. (	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mr:	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, M	female	26	0	0	STON/O2. 31	7.925		S
5	4	1	1	Futrelle, Mrs	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. Wi	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. J	male		0	0	330877	8.4583		Q

Calculate R2 Scores for different subsets of features.

Allows us to score the importance of different features

“Controlling for Y”

How much more predictive is the model if you additionally consider feature x over y?

R2([featurex, featurey]) - R2([featurey])

Explained Variation

$$1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Genetics explains 75% of human height.

Controlling for socio-economic status, age explains 80% of all cases.

…