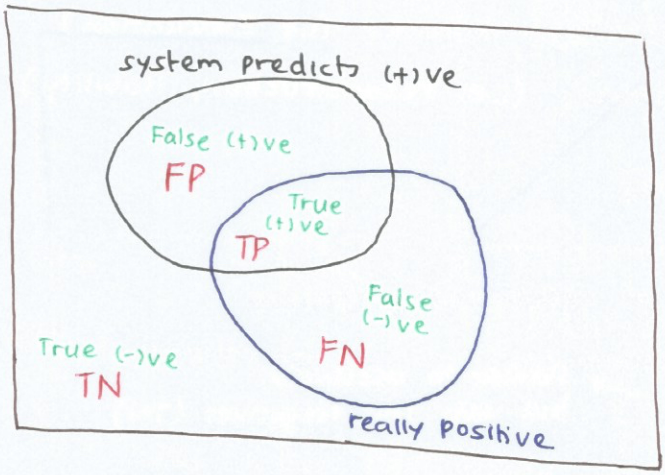
EVALUATION

→ Evaluation measure is a measure of how accurate a system is on a task

CLASSIFICATION (1).

all testing instances



Really positive?	Predict Positive? Yes	No
Yes	TP	FN
No	FP	TN

Confusion matrix for two-class classification

→ Want large TP and TN

→ Want small FP and FN

CLASSIFICATION ERROR

→ $\text{Classification Error} = \frac{\text{errors}}{\text{total}} = \frac{FP + FN}{TP + TN + FP + FN}$

→ $\text{Accuracy} = (1 - \text{Error}) = \frac{\text{correct}}{\text{total}} = \frac{TP + TN}{TP + TN + FP + FN}$

PROBLEM → Misleading if we have imbalanced classes

i.e. Predict whether an earthquake is about to happen (Rare event)

- Very good accuracy if always predict 'No'
- SOLN: Make FNs much more costly than FPs

SOLN

→ **False Alarm rate** — % of negatives we misclassified as positive

$= \frac{FP}{FP + TN}$

→ **Miss rate** — % of positives we misclassified as negative

$= \frac{FN}{TP + FN}$

→ **Recall** — % of positives we classified correctly ($1 - \text{miss rate} = \frac{TP}{TP + FN}$)

→ **Precision** — % of positives out of what we predicted was positive $= \frac{TP}{TP + FP}$

CLASSIFICATION COST & UTILITY

→ Sometimes we need a single-number evaluation measure in cases where we're doing something automated

1) Detection Cost — weighted avg. of FP and FN rates

$$\text{Cost} = C_{FP} \times FP + C_{FN} \times FN$$

i.e. Predicting earthquakes

↳ C_{FP} could be cost of preventive measures (i.e. Evacuation)

↳ C_{FN} could be cost of recovery (i.e. reconstruction, liability)

→ Used widely in event detections

2) F-measure — harmonic mean of recall and precision

$$F1 = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

Similar to accuracy, but w/o True Negatives (TN)

→ Used widely in information retrieval

THRESHOLDS IN CLASSIFICATION

→ If we have the ff. performance of 2 classifiers: A and B

	A	B
TP	50 %	100 %
FP	20 %	60 %

Which is better?

↳ Impossible to decide! A and B could be the exact same system operating at different thresholds

→ Many algorithms compute a confidence fn $f(x)$ and compare it to a threshold

i.e. We consider an email is spam if $f(x) > t$ and non-spam if $f(x) \leq t$

In Naive Bayes, $P(\text{spam} \mid x) > 0.5$

- Threshold t determines error rates

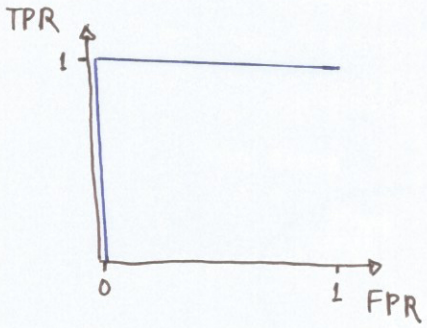
↳ False Positive Rate = $P(f(x) > t \mid \text{ham})$

True Positive Rate = $P(f(x) > t \mid \text{spam})$

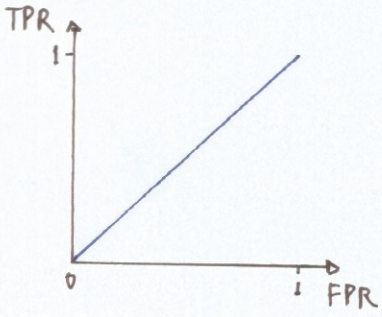
ROC CURVE

→ Receiver Operating Characteristic

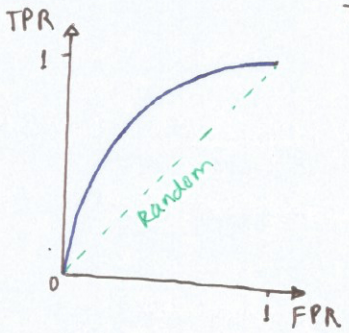
- A type of graph that tries to take threshold out of the equation for evaluation
- Instead of setting a particular threshold, we try ALL possible thresholds and plot graph of TPR vs. FPR as t varies from ∞ to $-\infty$ to show the performance of the system



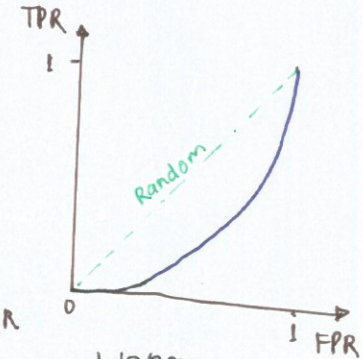
PERFECT performance



RANDOM Performance



BETTER THAN RANDOM



WORSE THAN RANDOM

- Area under ROC curve (AUC) is a popular alternative to accuracy

REGRESSION (2).

→ We want to measure how close we are to what we're trying to predict

→ Three ways to evaluate regression:

1) Mean squared Error (MSE)

→ Avg. squared deviation from truth

$$\text{Root MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - y_i\right)^2}$$

→ Very sensitive to single large errors (outliers) → large effect on our model

i.e. System A: predicts 99 prices exactly, and 1 price wrong by $10
System B: predicts all 100 prices wrong by $1
→ Both have the same MSE! (Due to squaring)

→ Very sensitive to the mean & scale of the prediction

- Getting the mean correctly is very important
- $$\text{Relative Squared Error} = \sqrt{\frac{\sum_{i=1}^{n}\left(f(x_i) - y_i\right)^2}{\sum_{i=1}^{n}\left(\mu_Y - y_i\right)^2}}$$

numerator: MSE of predictor
denominator: MSE when using the mean as a predictor

2) Mean Absolute Error (MAE)

→ less sensitive to outliers

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - Y_i|$$

→ The best 0^{th} order baseline is not the mean, but the median

→ Not differentiable and sensitive to the mean & scale

3) Correlation Coefficient

→ Insensitive to the mean / scale

$$\text{Correlation Coefficient} = \frac{n \sum_{i=1}^{n} (f(x_i) - \mu_i)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{n} (f(x_i) - \mu_i)^2 \cdot \sum_{i=1}^{n} (Y_i - \mu_Y)^2}}$$