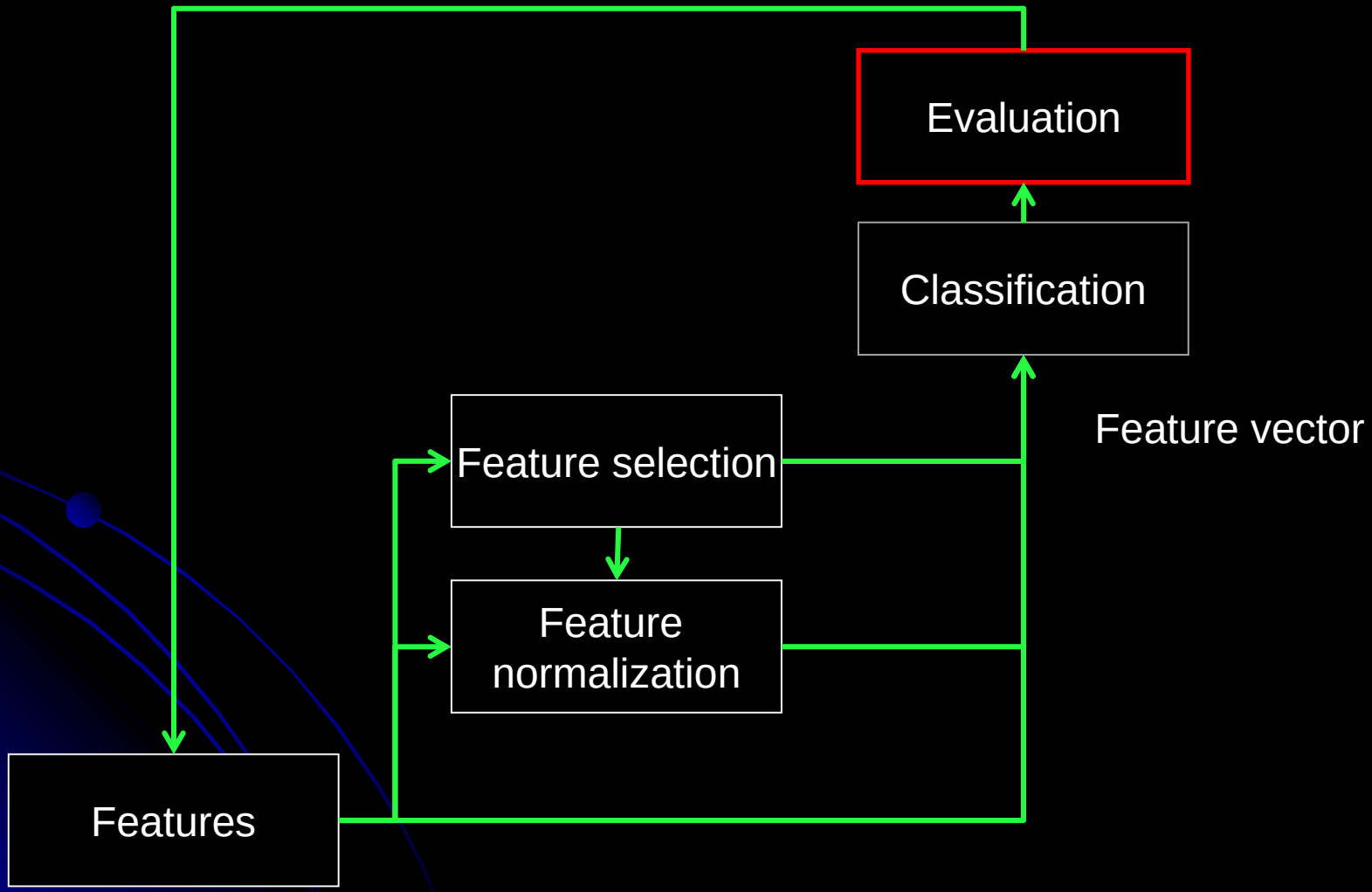
Machine learning in computer vision

Lesson 4

Classification pipeline



Evaluation of Results

How do you report classification error?

How certain are you about the error you claim?

How do you compare two algorithms?

How certain are you if you state one algorithm performs better than another?

Model Evaluation

Metrics for Performance Evaluation

- How to evaluate the performance of a model?

Methods for Performance Evaluation

- How to obtain reliable estimates?

Methods for Model Comparison

- How to compare the relative performance among competing models?

Model Evaluation

Metrics for Performance Evaluation

- How to evaluate the performance of a model?

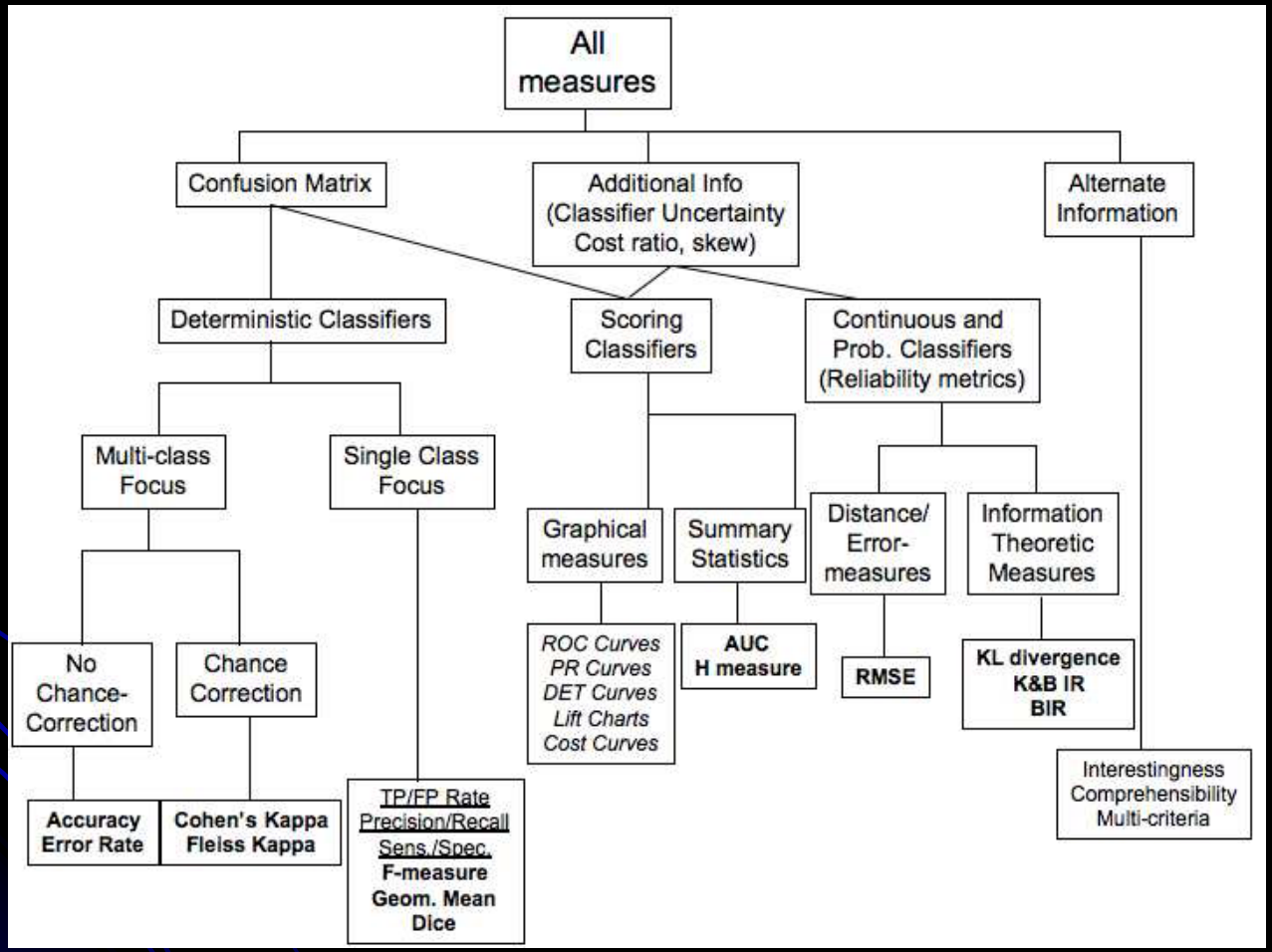
Methods for Performance Evaluation

- How to obtain reliable estimates?

Methods for Model Comparison

- How to compare the relative performance among competing models?

Performance measures



Nathalie Japkowicz & Mohak Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, 2011

Confusion matrix – binary classification

truth \ prediction	Positive	Negative	# of examples
Positive	TP (True Positive)	FN (False Negative)	P
Negative	FP (False Positive)	TN (True Negative)	N

TP: number of positive examples correctly predicted

FN: number of positive examples wrongly predicted as negative

FP: number of negative examples wrongly predicted as positive

TN: number of negative examples correctly predicted

Performance metrics

Percentage of correctly classified instances

$$Accuracy = \frac{TP+TN}{P+N}$$

Problem?

truth \ prediction	Positive	Negative	# of examples
Positive	0	1	1
Negative	0	99	99

Accuracy=99%

Aliases and other measures

t \ p	P	N
P	TP	FN
N	FP	TN

Accuracy = 1 – Error rate

Recall = True Positive Rate (TPR) = Hit rate = Sensitivity

Precision = Positive Predictive Value (PPV)

Negative Predictive Value (NPV) = TN/(TN+FN)

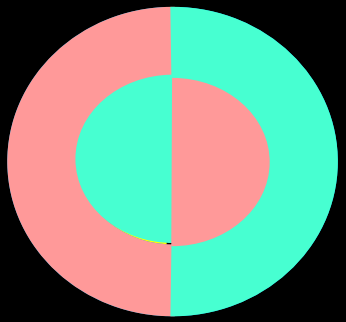
Fallout = False Positive Rate (FPR) = False Alarm rate = FP/(FP+TN)

Precision Recall curve

truth \ prediction	Positive	Negative	# of examples
Positive	TP (True Positive)	FN (False Negative)	P
Negative	FP (False Positive)	TN (True Negative)	N

The proportion of positives that are correctly identified as such

Recall = $\frac{TP}{P}$



The fraction of relevant instances among the selected ones

Precision = $\frac{TP}{TP+FP}$

Precision Recall curve

Example:

A method outputs the probability of belonging to class 1

A classifier assigns object to class 1 if outcome>threshold

Precision Recall curve

Example values:

Threshold	TP	FP	TN	FN
0	50	50	0	0
0,1	48	47	3	2
0,2	47	40	10	3
0,3	45	31	19	5
0,4	44	23	27	6
0,5	42	16	34	8
0,6	36	12	38	14
0,7	30	11	39	20
0,8	20	4	46	30
0,9	12	3	47	38
1	0	0	50	50

Precision and Recall values

Precision	Recall
0,5	1
0,505263158	0,96
0,540229885	0,94
0,592105263	0,9
0,656716418	0,88
0,724137931	0,84
0,75	0,72
0,731707317	0,6
0,833333333	0,4
0,8	0,24
#DIV/0!	0

Precision Recall curve

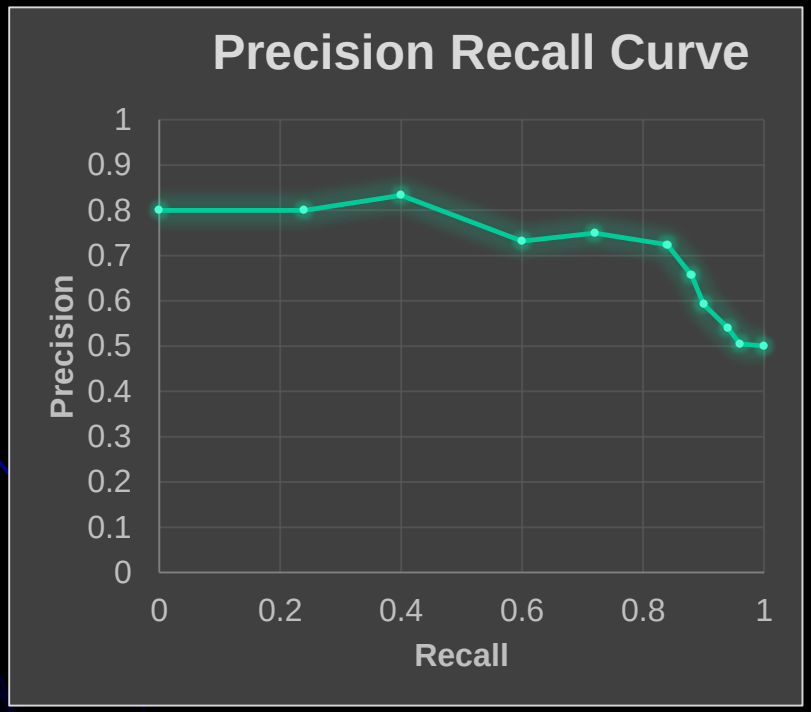
Precision and Recall values

Precision	Recall
0,5	1
0,505263158	0,96
0,540229885	0,94
0,592105263	0,9
0,656716418	0,88
0,724137931	0,84
0,75	0,72
0,731707317	0,6
0,833333333	0,4
0,8	0,24
0,8	0

Precision Recall curve

The end point of the precision-recall curve is always

Precision = P / (P + N), Recall = 1.0

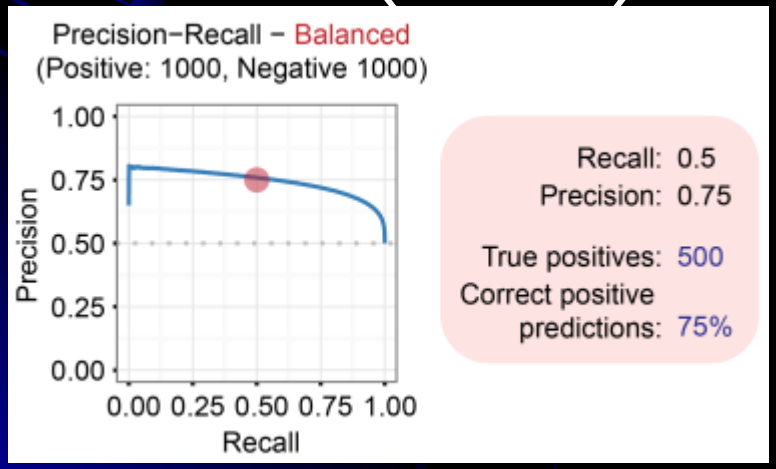


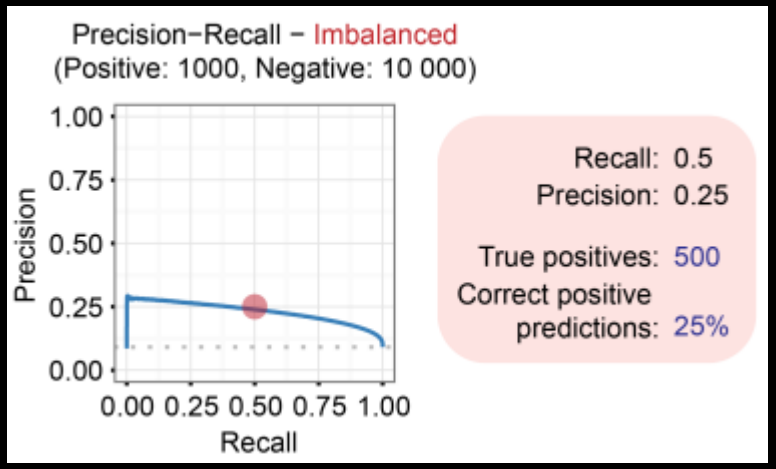
Precision Recall curve

Imbalanced data

	# of positives	# of negatives
Balanced	1000	1000
Imbalanced	1000	10 000

Baseline of the PR curve is given by the ratio P / (P + N)



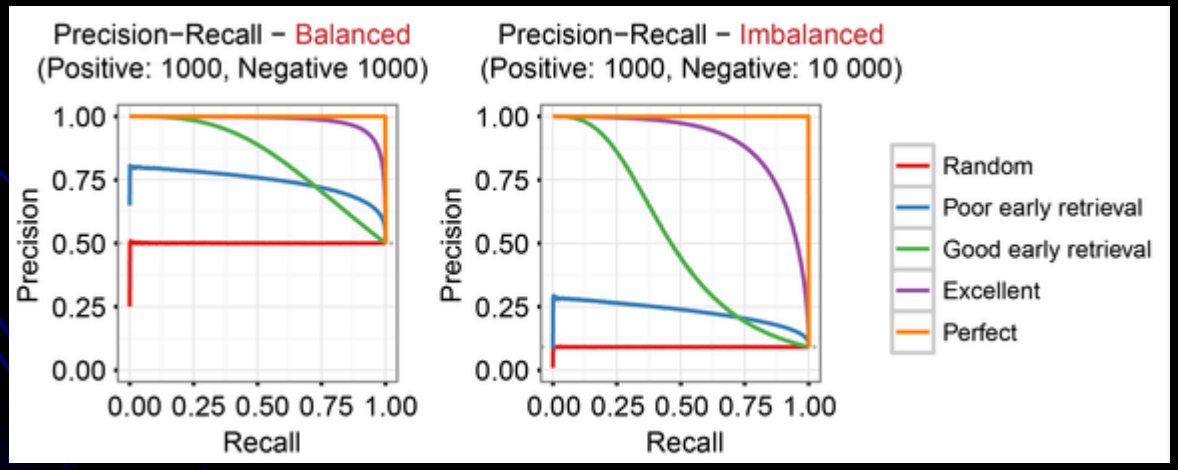


https://classeval.wordpress.com/

Precision Recall curve shape

Random classifier
Assigns object to class 1 with probability
P / (P + N)

Perfect classifier
Assigns objects correctly



https://classeval.wordpress.com/

Precision Recall curve

Average precision:

precision averaged across all values of recall between 0 and 1

$\mathrm{AP} = \int_0^1 p(r)\mathrm{d}r$

Approximated average precision:

sum over the precisions at every possible threshold value t_i, multiplied by the change in recall

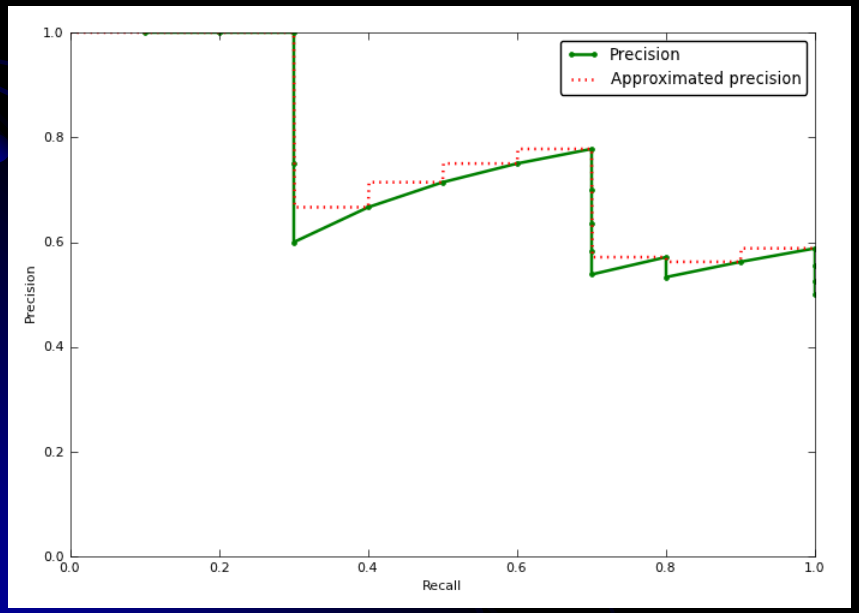
$\mathrm{AP} = \sum_{i=1}^{T} p(t_i).\Delta r(t_i)$

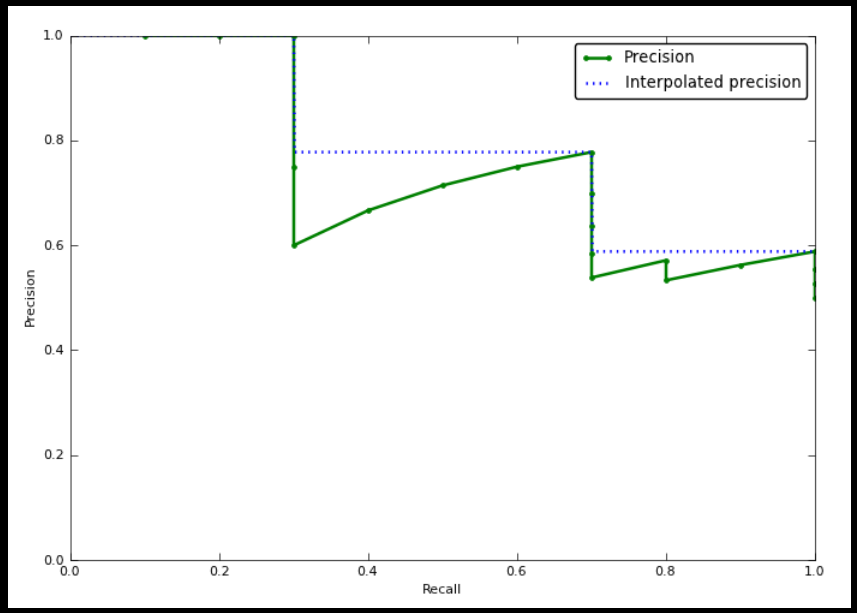
Precision Recall curve

interpolated average precision:

uses the maximum precision observed across all thresholds with higher recall

$$\mathrm{AP} = \sum_{i=1}^{T} \max_{r(t) \geq r(t_i)} p(t) \,.\, \Delta r(t_i)$$





https://sanchom.wordpress.com/2011/09/01/precision-recall/

F score

Precision and recall combined into a single measure, computed at a specific detection threshold

$$F_\beta(t) = \frac{(1+\beta^2)p(t)r(t)}{r(t)+\beta^2 p(t)}$$

$\beta = 1$: harmonic mean

If we want to create a balanced classification model with the optimal balance of recall and precision, then we try to maximize the F_1 score

Performance metrics

truth \ prediction	Positive	Negative	# of examples
Positive	TP (True Positive)	FN (False Negative)	P
Negative	FP (False Positive)	TN (True Negative)	N

The proportion of positives that are correctly identified as such

$$\text{TPR} = \frac{TP}{P}$$

The proportion of negatives identified as positives

$$\text{FPR} = \frac{FP}{N}$$

ROC curve

Receiver Operating Characteristics (ROC)

Example:

A method outputs the probability of belonging to class 1

A classifier assigns object to class 1 if outcome>threshold

ROC curve

Example values:

Threshold	TP	FP	TN	FN
0	50	50	0	0
0,1	48	47	3	2
0,2	47	40	10	3
0,3	45	31	19	5
0,4	44	23	27	6
0,5	42	16	34	8
0,6	36	12	38	14
0,7	30	11	39	20
0,8	20	4	46	30
0,9	12	3	47	38
1	0	0	50	50

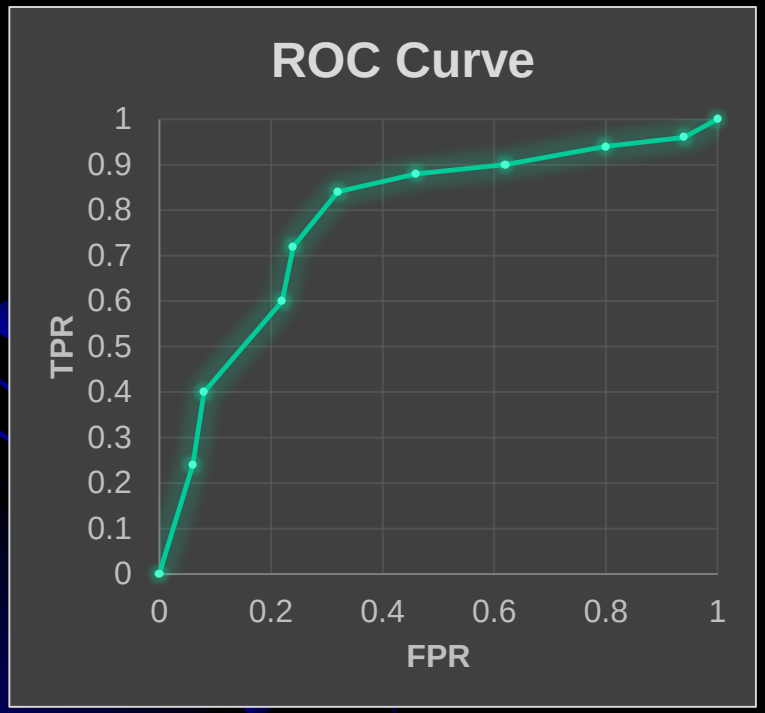
TPR and FPR values

TPR	FPR
1	1
0,96	0,94
0,94	0,8
0,9	0,62
0,88	0,46
0,84	0,32
0,72	0,24
0,6	0,22
0,4	0,08
0,24	0,06
0	0

ROC curve

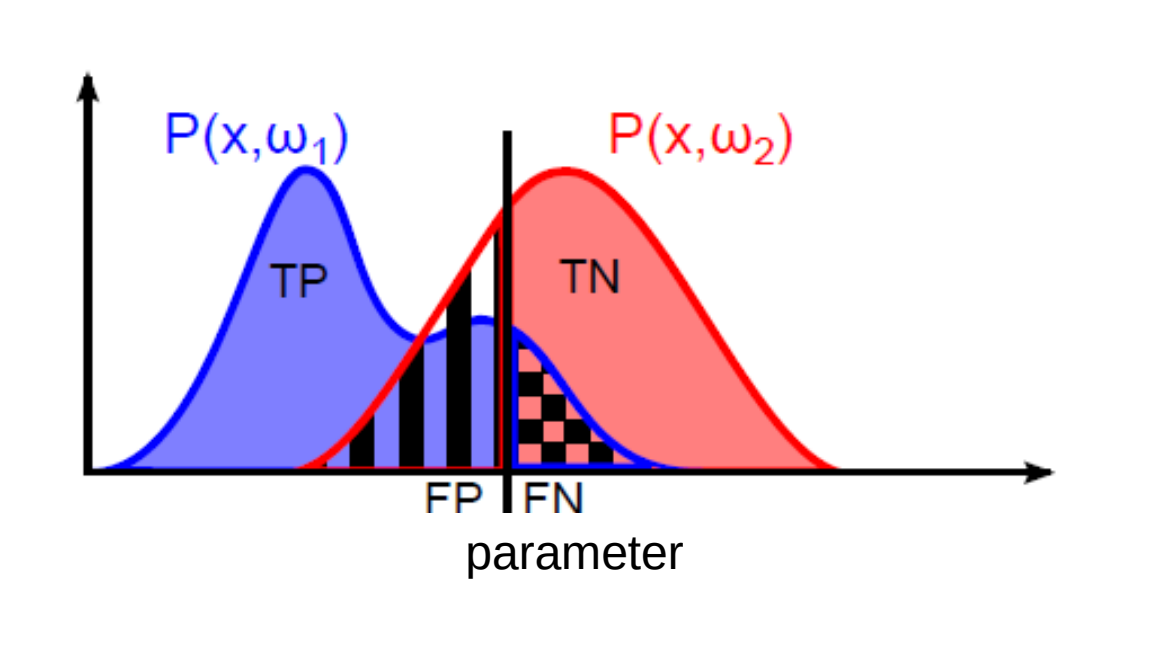
TPR and FPR values

TPR	FPR
1	1
0,96	0,94
0,94	0,8
0,9	0,62
0,88	0,46
0,84	0,32
0,72	0,24
0,6	0,22
0,4	0,08
0,24	0,06
0	0

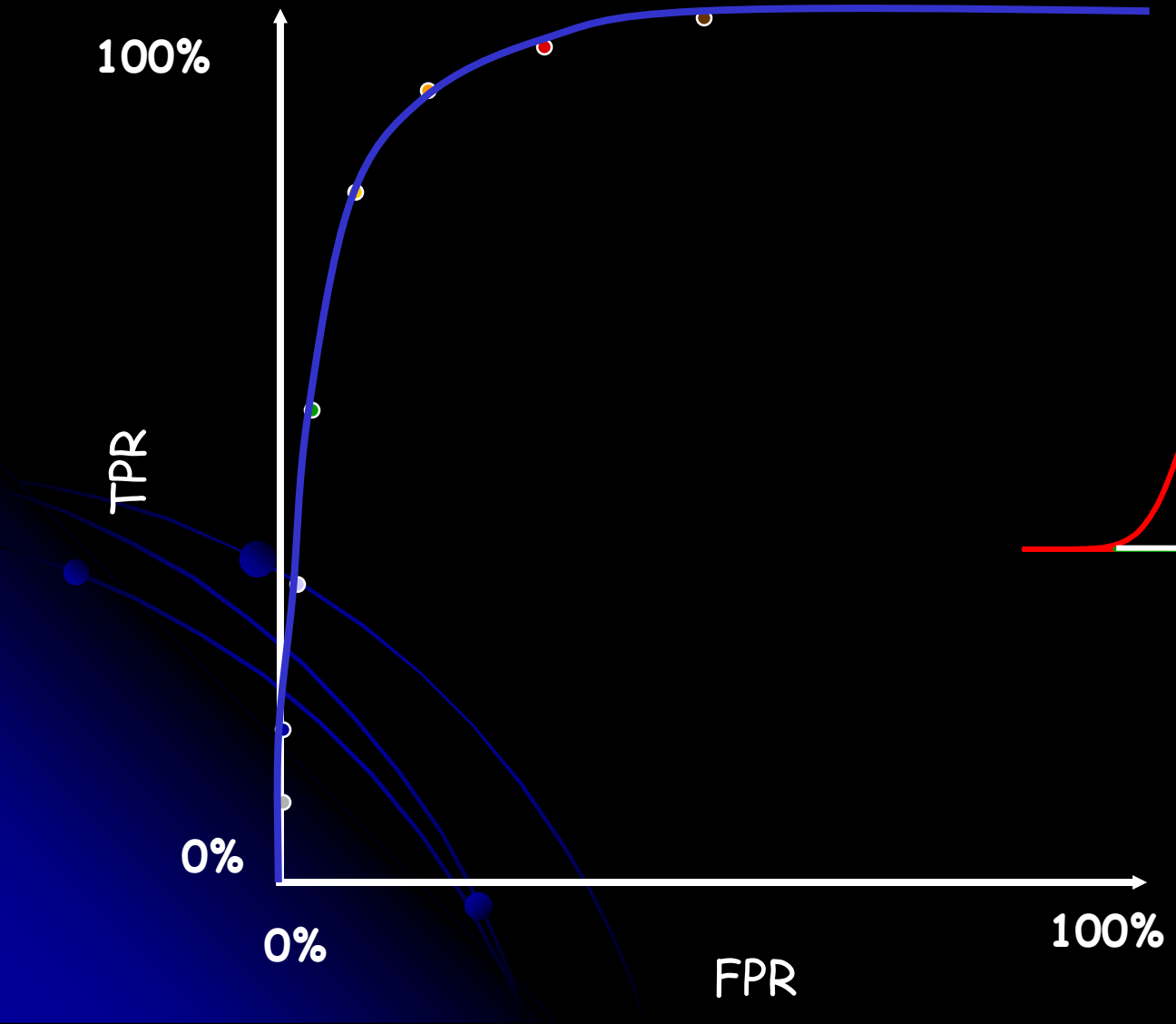


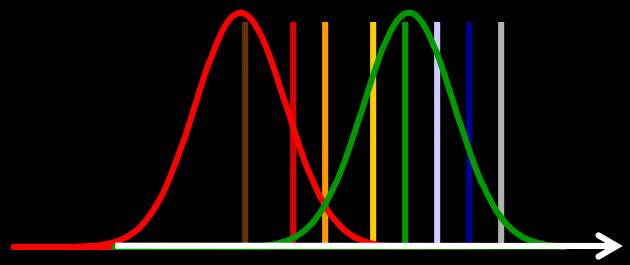
ROC curve

Uses all entries of the confusion matrix

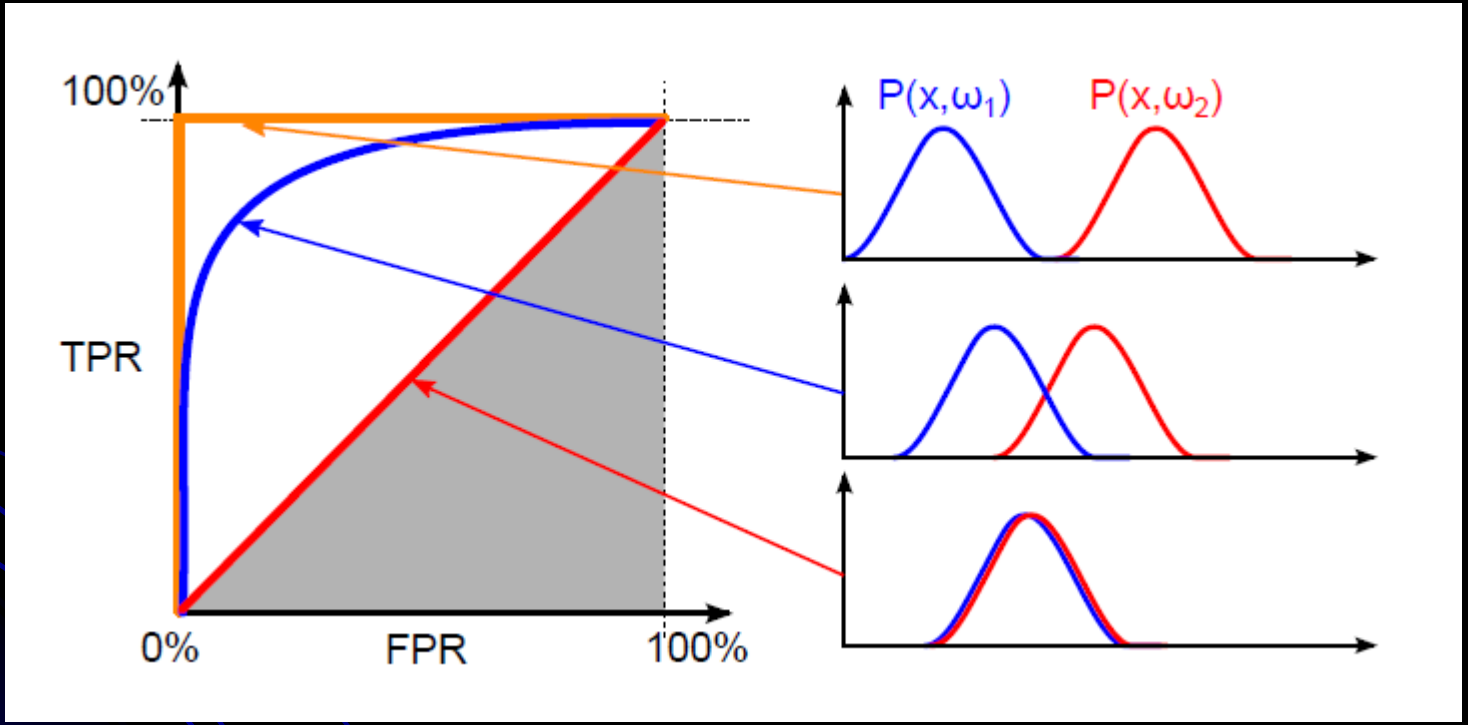


ROC curve



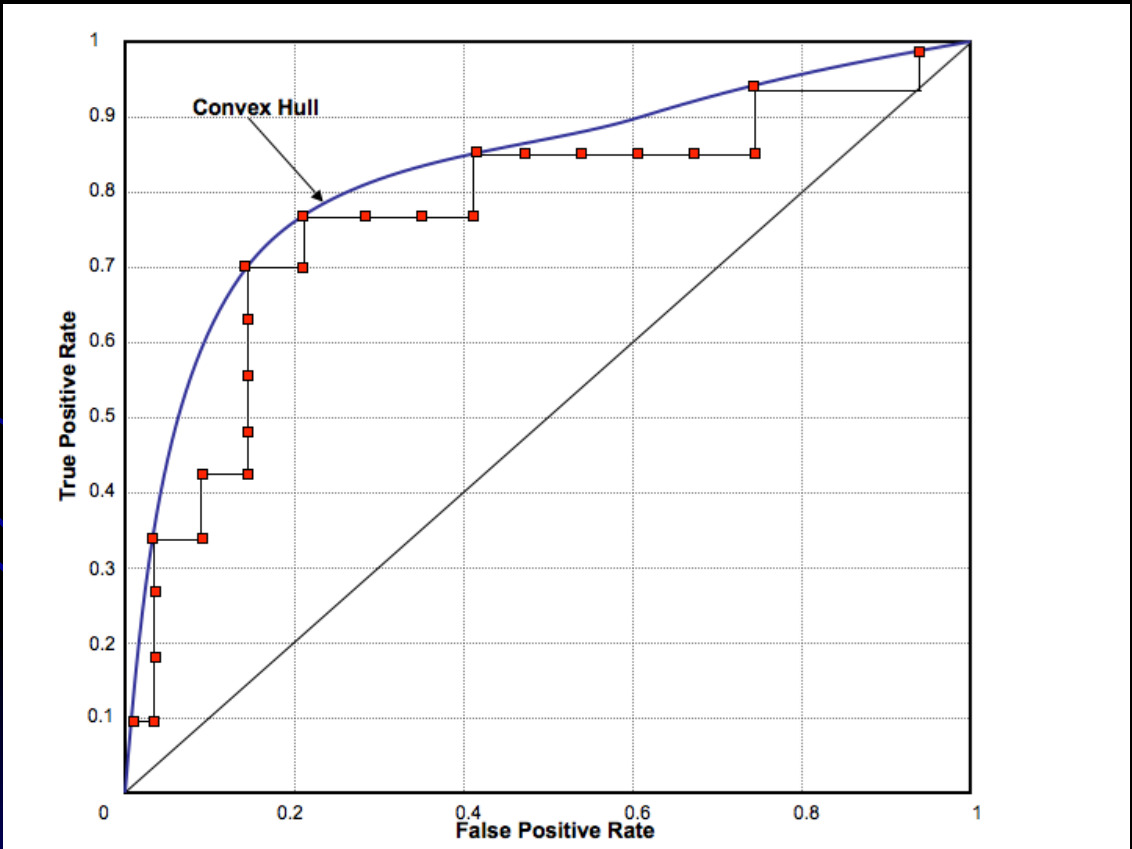


ROC curve shape



ROC curve

The set of points on ROC curve that are not suboptimal forms the ROC *convex hull*



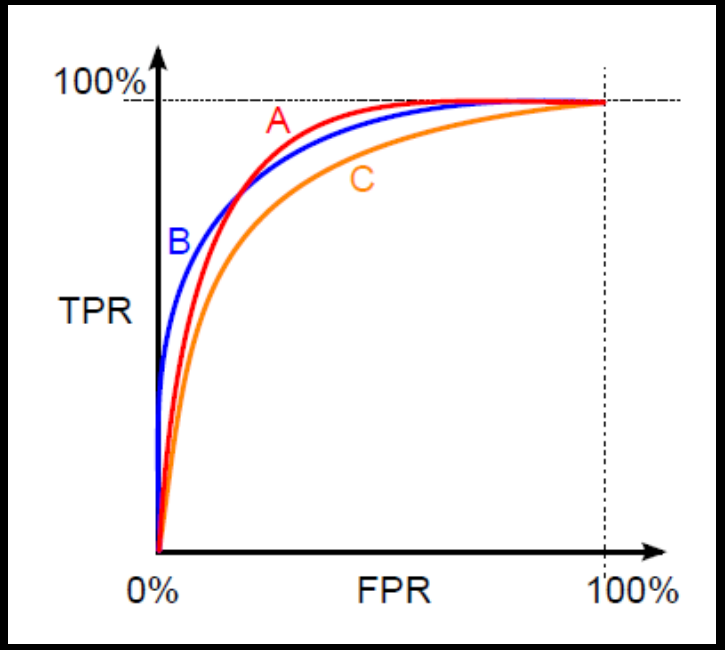
ROC curve

AUC – area under curve

equal to the probability that a random positive example will be ranked above a random negative example

Area Properties

- 1.0 - Perfect prediction
- .9 - Excellent
- .7 - Mediocre
- .5 - Random

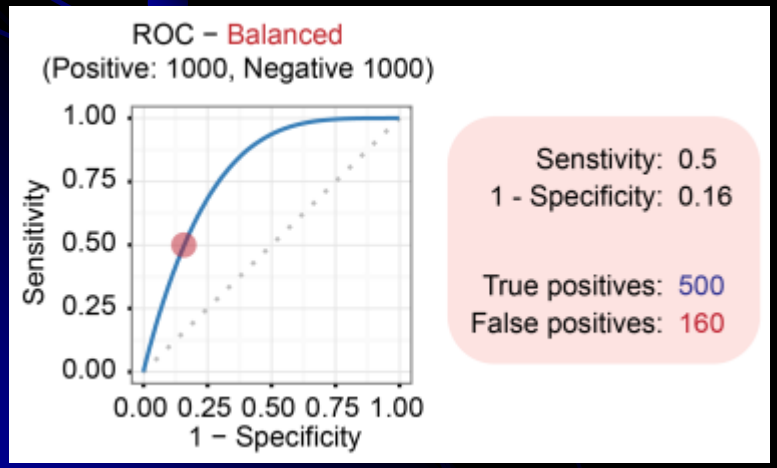


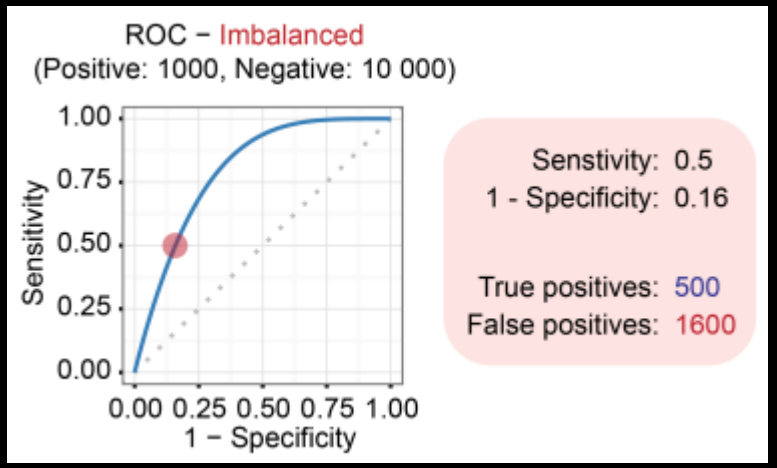
ROC curve

Imbalanced data

	# of positives	# of negatives
Balanced	1000	1000
Imbalanced	1000	10 000

ROC curves appears to be identical under balanced and imbalanced cases





https://classeval.wordpress.com/

ROC and PR curves

ROC curves should be used when there are roughly equal numbers of observations for each class.

Precision-Recall curves should be used when there is a moderate to large class imbalance.

Loss function

$L: Y \times Y \rightarrow R^{+}$ indicating the penalty for an incorrect prediction

$\hat{y} = f(\mathbf{x})$

$L(\hat{y}, y)$ – loss for prediction of $\hat{y}$ instead of y

$R_{\hat{y}}(\mathbf{x}) = \sum_{y \epsilon C} L(\hat{y}, y) P(y|\mathbf{x})$ – expected risk from classifying **x** as class $\hat{y}$

$R = \sum_{\hat{y} \epsilon C} \int_{\mathcal{R}_{\hat{y}}} R_{\hat{y}}(\mathbf{x})\, p(\mathbf{x}) \mathrm{d}\mathbf{x}$

$= \sum_{y \epsilon C} \sum_{\hat{y} \epsilon C} \int_{\mathcal{R}_{\hat{y}}} L(\hat{y}, y) P(y|\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x}$ – expected risk from using the classifier

Loss function

zero-one loss: standard loss function in classification

$$L(\hat{y}, y) = 1 - \delta_{\hat{y},y}$$

Classifying an object as face or non-face is almost equally costly

truth \ prediction	Face	Non-face
Face	0	1
Non-face	1	0

Loss function

non-symmetric losses:

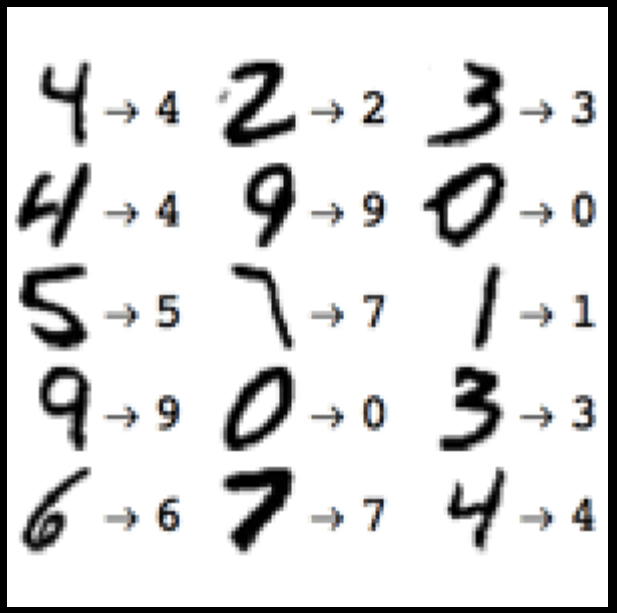
$$L(\hat{y}, y) > L(y, \hat{y})$$

does not have cancer, but is classified as having it ⇨ distress, plus unnecessary further investigation

does have cancer, but is classified as not having it ⇨ no treatment, premature death

truth \ prediction	cancer	healthy
cancer	0	1000
healthy	1	0

Multiple classes



Multiple classes

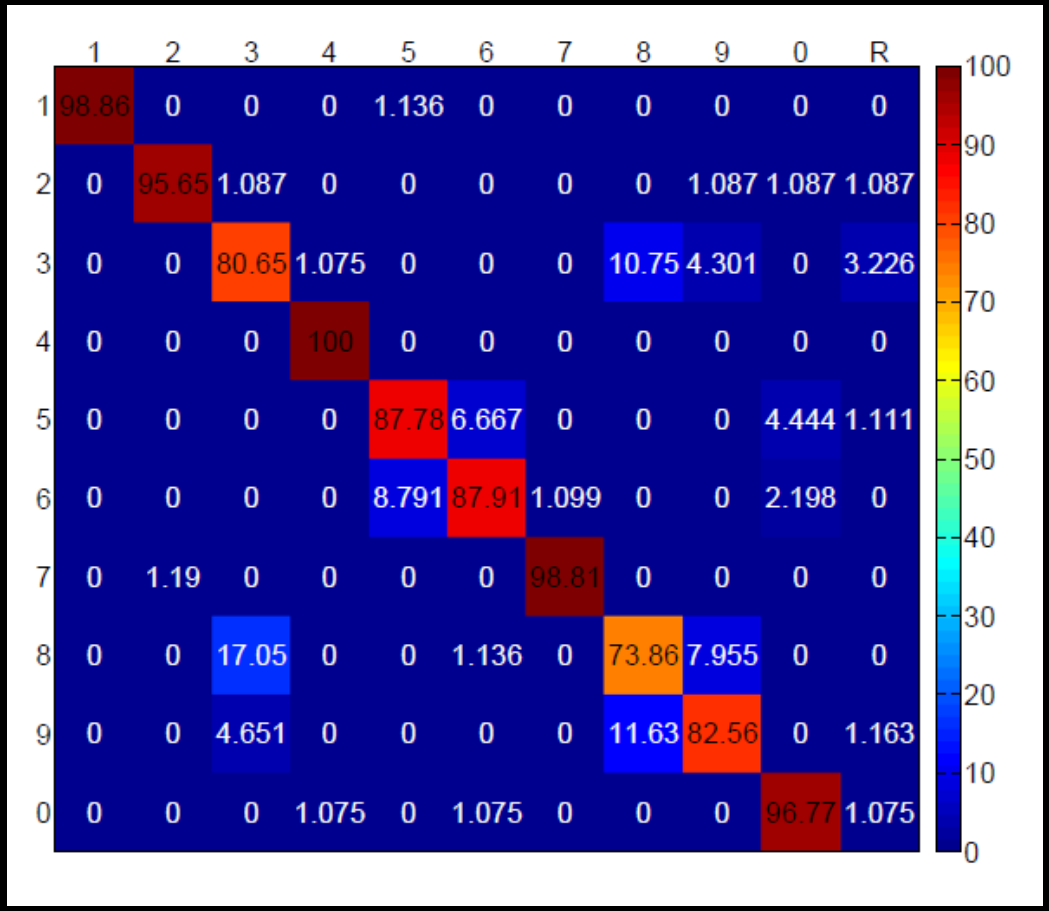
number	classification										
	1	2	3	4	5	6	7	8	9	0	R
1	87	0	0	0	1	0	0	0	0	0	0
2	0	88	1	0	0	0	0	0	1	1	1
3	0	0	75	1	0	0	0	10	4	0	3
4	0	0	0	79	0	0	0	0	0	0	0
5	0	0	0	0	79	6	0	0	0	4	1
6	0	0	0	0	8	80	1	0	0	2	0
7	0	1	0	0	0	0	83	0	0	0	0
8	0	0	15	0	0	1	0	65	7	0	0
9	0	0	4	0	0	0	0	10	71	0	1
0	0	0	0	1	0	1	0	0	0	90	1

Special class *R* – *reject*

Multiple classes

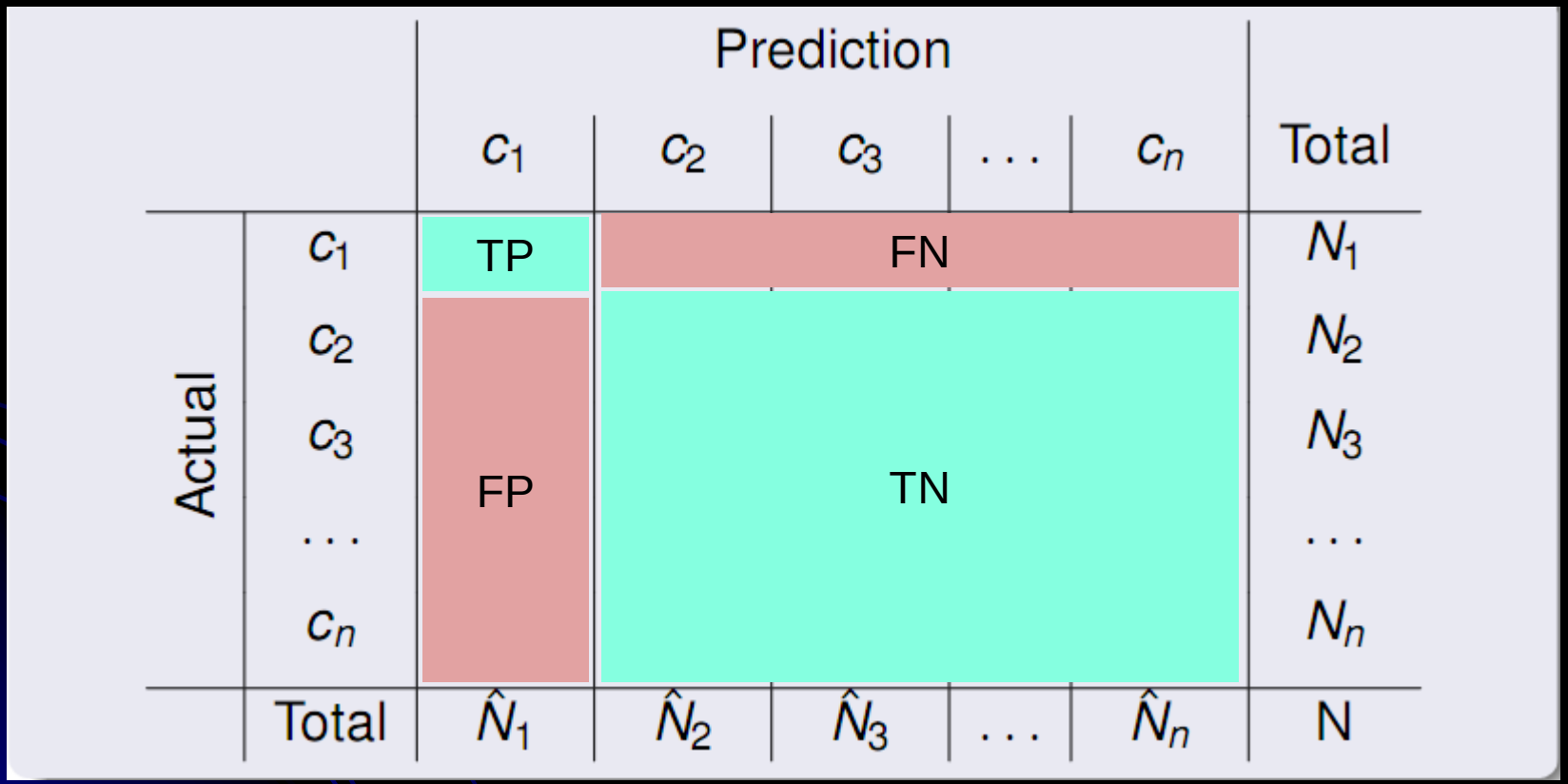
Heatmap visualization

Class percentage



Multiple classes

One vs. all



http://www.icmla-conference.org/icmla10/CFP_Tutorial_files/jose.pdf

http://rali.iro.umontreal.ca/rali/sites/default/files/publis/SokolovaLapalme-JIPM09.pdf

Multiple classes

- Macro average
  - Precision, recall in each class – compute (weighted) mean precision and recall
- Micro average
  - Sum TPs, FPs and FNs of all classes, compute precision and recall

Sokolova, Marina & Lapalme, Guy. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management. 45. 427-437.

Model Evaluation

Metrics for Performance Evaluation

- How to evaluate the performance of a model?

Methods for Performance Evaluation

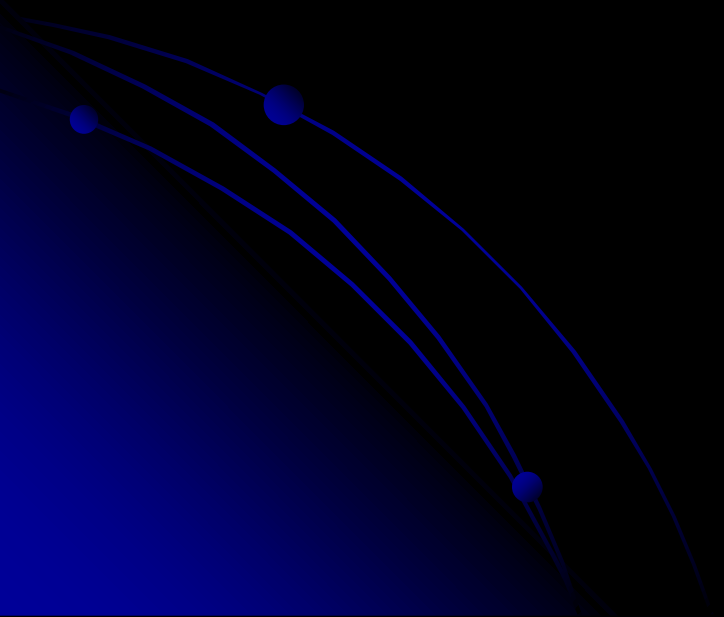
- How to obtain reliable estimates?

Methods for Model Comparison

- How to compare the relative performance among competing models?

Methods for Performance Evaluation

The resulting *model* is also called the *hypothesis*.



Sample and True Error

$h(x)$ is the hypothesis under investigation

$error(h|X) = \frac{1}{N}\sum_{i=1}^{N}[h(\mathbf{x}_i) \neq y_i]$
is the error on the sample X

- $error(h|P) = \int [h(x) \neq t(x)]p(x)\mathrm{dx}$
is the true error on the unseen data sampled from the distribution $P(x)$, where $t(x)$ is the true hypothesis

Sample and True Error

We wish to know $error(h|P)$, but we can only compute $error(h|X)$

How good is the estimate of $error(h|P)$ provided by $error(h|X)$?

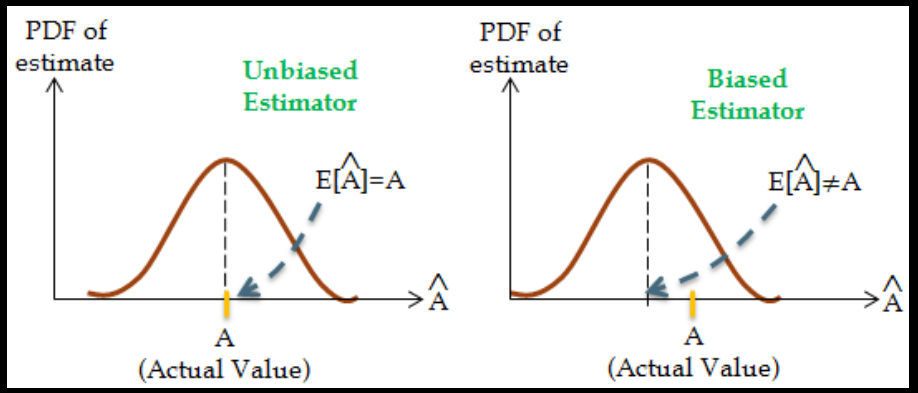
$error(h|X)$ is a random variable (outcome of an experiment)

Problems Estimating Error

Bias: If *X* is training set, $error(h|X)$ is optimistically biased

$$bias = E\big(error(h|X)\big) - error(h|P)$$

For unbiased estimate, *h* and *X* must be chosen independently



Variance: Even with unbiased *X*, $error(h|X)$ may still vary from $error(h|P)$

Binomial Experiment

experiment consists of n identical and independent trials

each trial results in one of two outcomes: success or failure

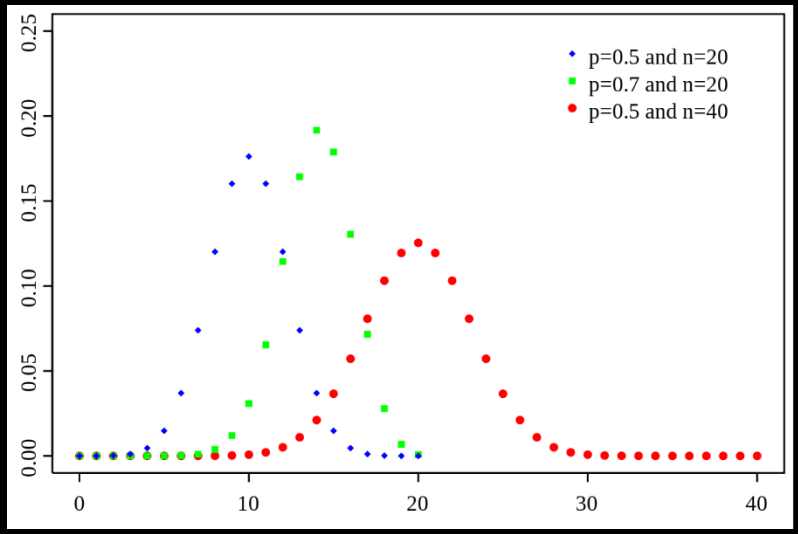
P(success) = p

P(failure) = 1 – p

The random variable of interest, X, is the number of successes in the n trials.

X has a binomial distribution with parameters n and p

$X \sim B(n,p)$



https://commons.wikimedia.org/wiki/File:Binomial_distribution_pmf.svg

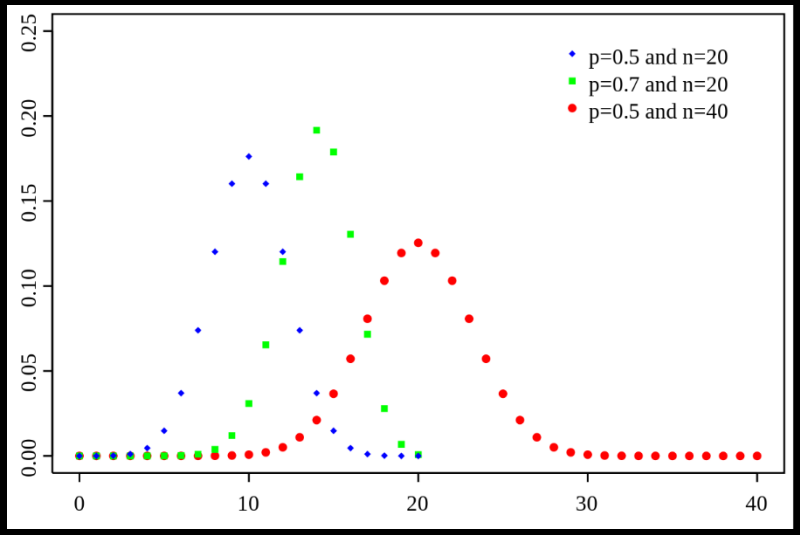
Binomial Distribution

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

Properties:

$$E[X] = np$$

$$V[X] = \sigma^2 = np(1-p)$$



https://commons.wikimedia.org/wiki/File:Binomial_distribution_pmf.svg

Classification as Binomial Experiment

Classification:

A fixed number of trials: n

Only two outcomes ("success" == $h(\mathbf{x}_i) \neq y_i$)

Probability of success in one trial: $p = error(h|P)$

Each trial is independent

$$error(h|X) = \frac{r}{n}$$

$$E\big(error(h|X)\big) = E\left(\frac{r}{n}\right) = \frac{E(r)}{n} = \frac{np}{n} = p = error(h|P)$$

$\rightarrow$ an unbiased estimator