Supervised learning: classification problems

Filippo Biscarini (CNR, Milan, Italy)

filippo.biscarini@cnr.it

Classification problems

Classification problems

- the response variable **y** is **qualitative**
- e.g.: coat colour, type of rice (Tropical japonica, Indica, Temperate japonica, Aromatic, Aus)
- **y** = **label** (a.k.a. dependent variable)
- **X** = matrix of **features** (continuous, categorical)

Classification problems

- **y** = **label** (a.k.a. dependent variable)
- **X** = matrix of **features** (continuous, categorical)
- we don't model the response (**y**) directly, rather its **probability**: **P(y=k|X)**
- probabilities lie in [0,1] (not +/- infinity)

Classification problems

classifier:

- K classes ($k \in K$)

probabilities

classifier

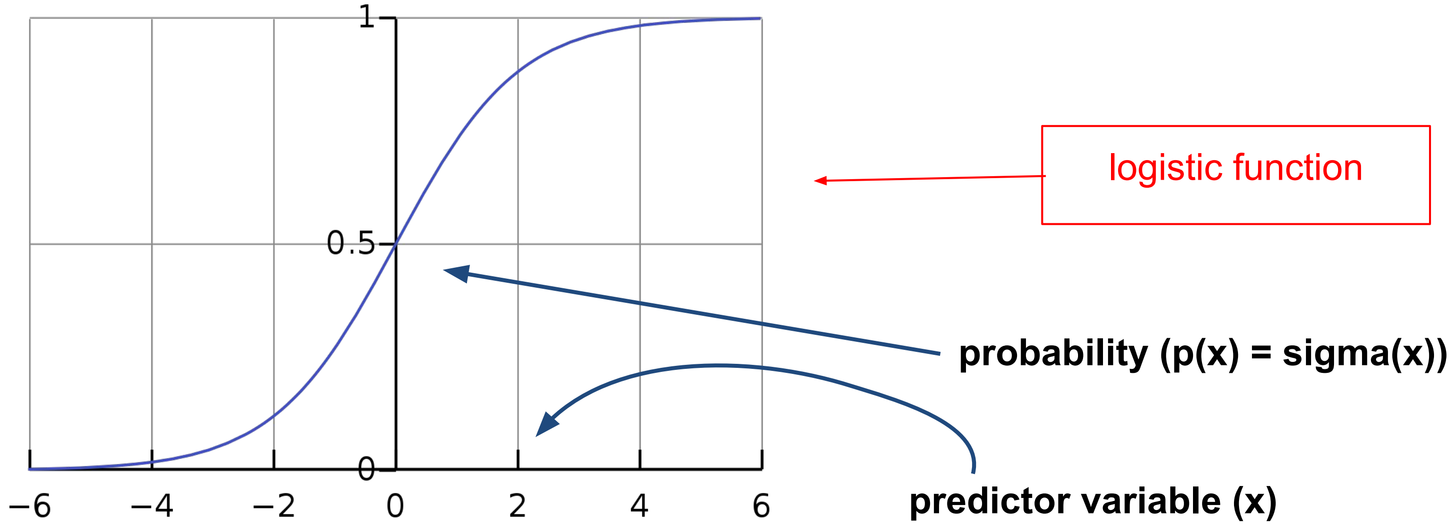
$$p_k(x) = Pr(y = k | X = x) = f(x)$$

$$C(x) = k, \text{ if } p_k(x) = max\{p_1(x), p_2(x), \dots, p_K(x)\}$$

Binary classification problems

- “special” classification case → only two classes
- binary traits (e.g. cases/controls, resistant/susceptible, high/low, 0/1 etc.)
  - can you think of other examples?
- no need to model the probability of the two classes: one suffices → $P(y=1|x) = f(x)$

Binary classification problems



$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{1}{1+\frac{1}{e^x}} = \frac{e^x}{1+e^x}$$

Logistic regression

- the logistic function is the basis for **logistic regression**
- **P(y=1|x)** [also p(x)]
- P(y=1|z) → **Z = $\boldsymbol{\beta_0 + \beta_1}$x** (linear combination of variables)

$$p(y=1|x) = \sigma(z) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

we see here the familiar **model coefficients** to be estimated and then used for predictions

Logistic regression

- a little bit of algebra:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \longrightarrow \frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$$

odds

Logistic regression

- a little bit of algebra:

odds

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \longrightarrow \frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$$

log(odds): logit

$$log\left(\frac{p(x)}{1-p(x)}\right) = logit(p(x)) = \beta_0 + \beta_1 x$$

Logistic regression

- the **logit function** (log(odds)) is the **link function** between a linear expression of X and the probabilities of Y
- linear X expression (β_0+β_1x) → logit scale (continuous)
- logistic function: converts values on the logit scale back to probabilities

$$\begin{cases} logit(p(x)) = \beta_0 + \beta_1 x \\ \sigma(\beta_0 + \beta_1 x) = p(x) \end{cases}$$

our objective!

Logistic regression - recap

1. the **logistic function** allows us to **model probabilities** in [0,1] as **functions of variables** (features)
2. we need to **transform** the **non-linear logistic expression** to a manageable **linear expression** →the **logit link** function
3. finally, we use again the **logistic function** to **convert** unbounded results on the **logit scale** to **probabilities** (of belonging to a class given the variables/features)

Estimating the coefficients

how do we obtain the model coefficients β?

- similarly to linear regression, we need to define a **cost function** and then minimise it

observations	**predictions**
y	$\hat{y} = \sigma(\beta_0 + \beta_1 x)$

difference between observed and predicted values → **LEAST SQUARES?**

Estimating the coefficients

how do we obtain the model coefficients β?

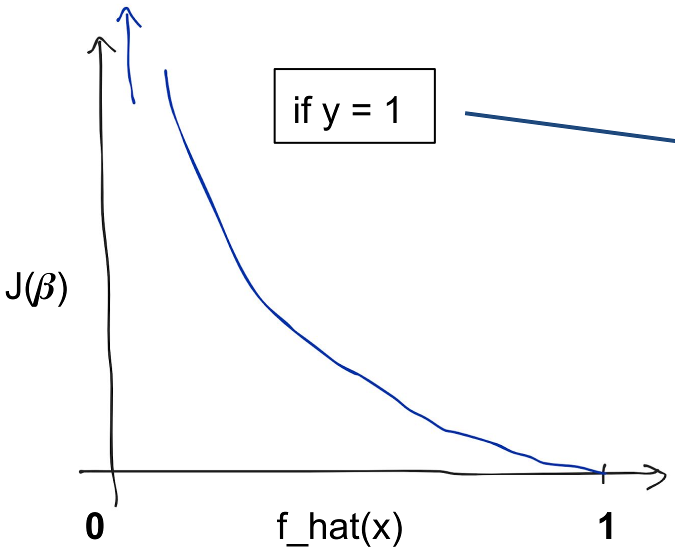
- similarly to linear regression, we need to define a **cost function** and then minimise it

$$J(\beta) = \text{Cost}(\hat{y}, y) = \begin{cases} -log(\hat{y}) & \text{if} y = 1 \\ -log(1-\hat{y}) & \text{if} y = 0 \end{cases}$$

$$J(\beta) = \text{Cost}(\hat{y}, y) = -\left(y \cdot log(\hat{y}) + (1-y) \cdot log(1-\hat{y})\right)$$

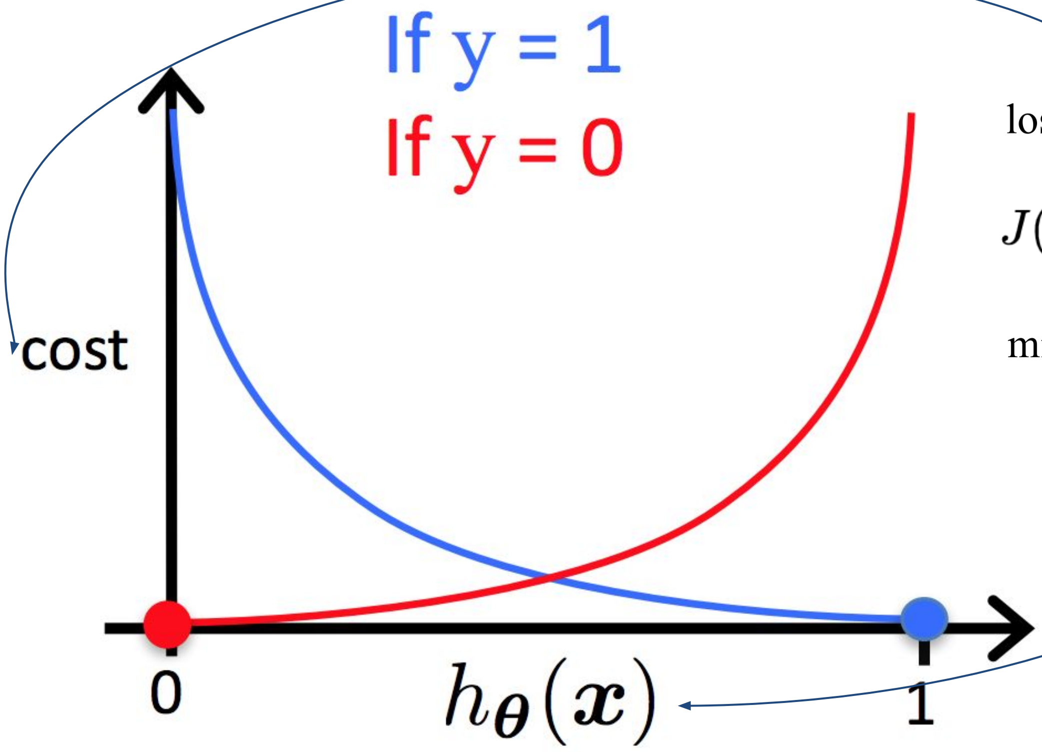
Cost function for logistic regression

$$J(\beta) = \text{Cost}(\hat{y}, y) = -(y \cdot log(\hat{y}) + (1 - y) \cdot log(1 - \hat{y}))$$



- if y_hat = 1, cost = 0
- if y_hat → 0 (but y = 1), cost → infinity
- the opposite holds if y = 0

Loss function for logistic regression



$$\text{loss}\,(\hat{y}, y) = J(\beta)$$

$$J(\beta) = -\left(y \cdot \log\left(\hat{y}\right) + (1-y) \cdot \log\left(1-y\right)\right)$$

$$\underset{\beta}{\text{minimize}}\, J(\beta)$$

From: https://datascience.stackexchange.com/questions/40982/logistic-regression-cost-function

Minimising the cost function

- the defined cost function is convex
- can be minimised by **gradient descent**
- machine learning perspective: gradient descent is a general algorithm to solve models
- alternatively:
    - maximum likelihood
    - non-linear least squares

Binary classification: model evaluation

- the most common metric to measure the performance of a binary classifier is the **error rate**:

$$\frac{1}{n}\sum_{i=1}^{n} I(y \neq \hat{y})$$

Confusion matrix

		True observation	
		1	0
Prediction	1	TP	FP
	0	FN	TN

Not only total error rate!

- **FPR** = FP/(FP+TN)
- **FNR** = FN/(FN+TP)
- TER = (FN+FP)/(FN+FP+TN+TP)

Introducing the dataset

Genetic variants for cleft lip in dogs

binary phenotypes: **cleft lip** (presence/absence)

PLOS | GENETICS

RESEARCH ARTICLE

Genome-Wide Association Studies in Dogs and Humans Identify *ADAMTS20* as a Risk Variant for Cleft Lip and Palate

Zena T. Wolf[1☯], Harrison A. Brand[2,3☯¤a], John R. Shaffer[3☯], Elizabeth J. Leslie[2], Boaz Arzi[4], Cali E. Willet[5], Timothy C. Cox[6,7,8], Toby McHenry[2], Nicole Narayan[9], Eleanor Feingold[3], Xioajing Wang[2¤b], Saundra Sliskovic[1], Nili Karmi[1], Noa Safra[1], Carla Sanchez[2], Frederic W. B. Deleyiannis[10], Jeffrey C. Murray[11], Claire M. Wade[5], Mary L. Marazita[2,12‡*], Danika L. Bannasch[1‡*]

Genetic variants for cleft lip in dogs

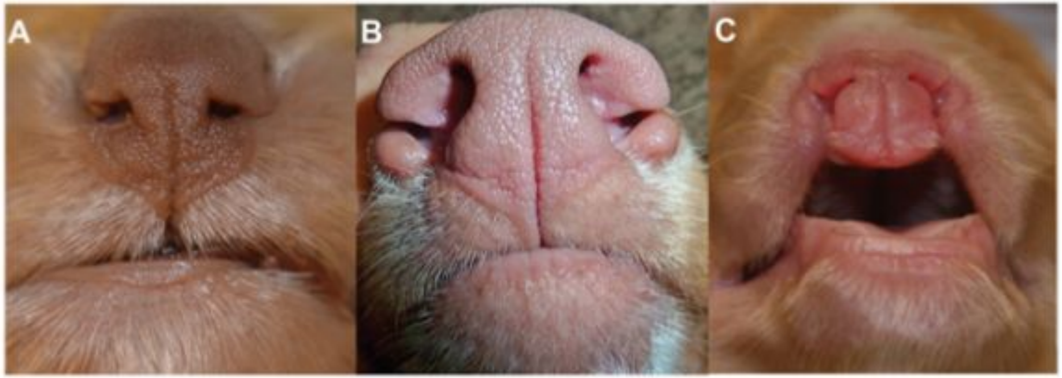
binary phenotype: **cleft lip** (presence/absence)

- Nova Scotia Duck Tolling Retriever (NSDTR)
- 125 dogs:
    - 13 cases
    - 112 controls

Genetic variants for cleft lip in dogs

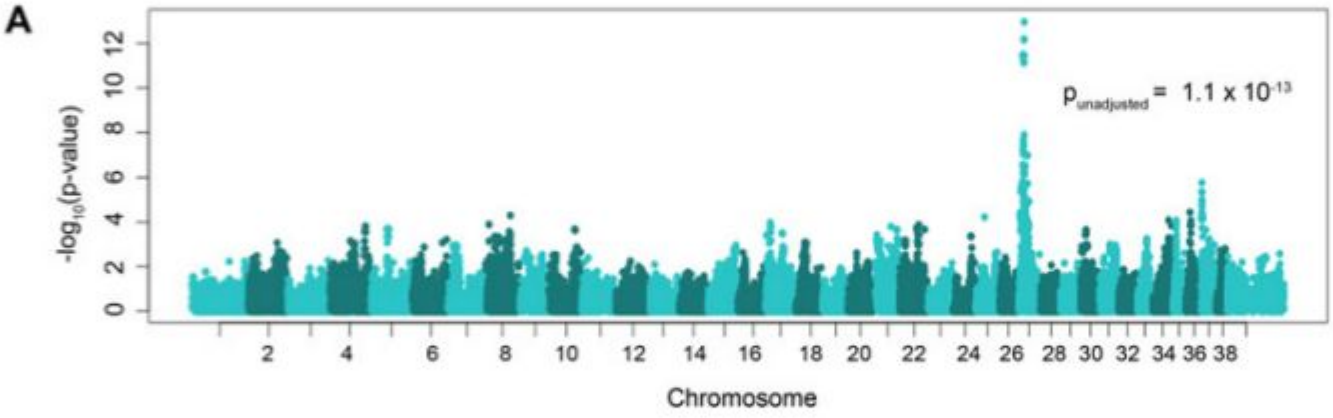
binary phenotypes: **cleft lip** (presence/absence)

- Nova Scotia Duck Tolling Retriever (NSDTR)
- 125 dogs:
    - 13 cases
    - 112 controls



Genetic variants for cleft lip in dogs

binary phenotypes: **cleft lip** (presence/absence)



39 chromosomes

Strong signal of association on chromosome 27

Logistic regression

- demonstration 4.1
- Exercise 4.1

→ 4.classification.ipynb

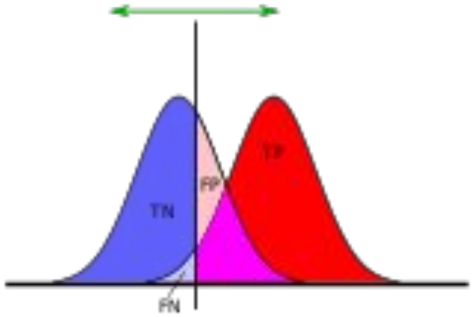
ROC curves

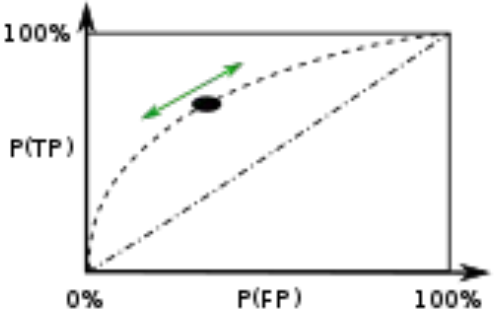
Binary classification

		True observation	
		1	0
Prediction	1	TP	FP
	0	FN	TN

- classify observations in **two categories** (1/0)
- however, predictions are usually **probabilities** (P(y=1|x))
- different ***cut-offs*** (e.g. 0.5 or 0.8 0r 0.3) will give different results

ROC curves

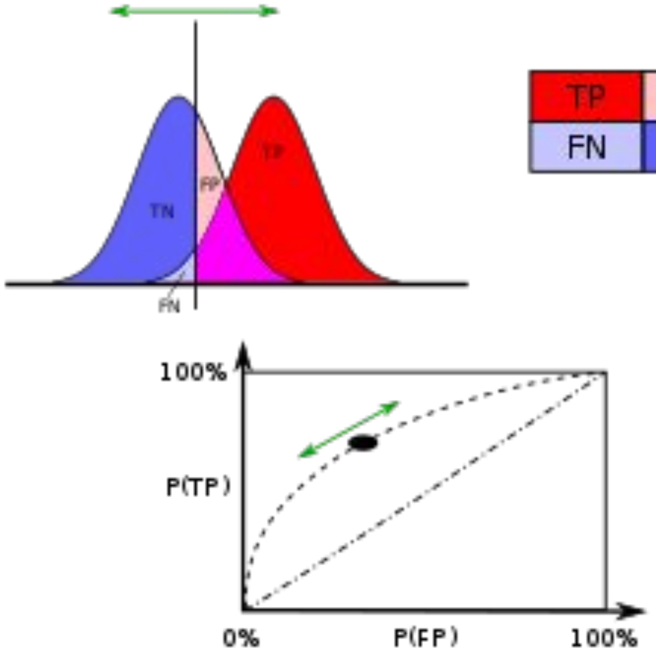


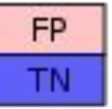


- Relationship between TPR and FPR
- The diagonal is chance classification (no predictive ability)

Source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

ROC curves





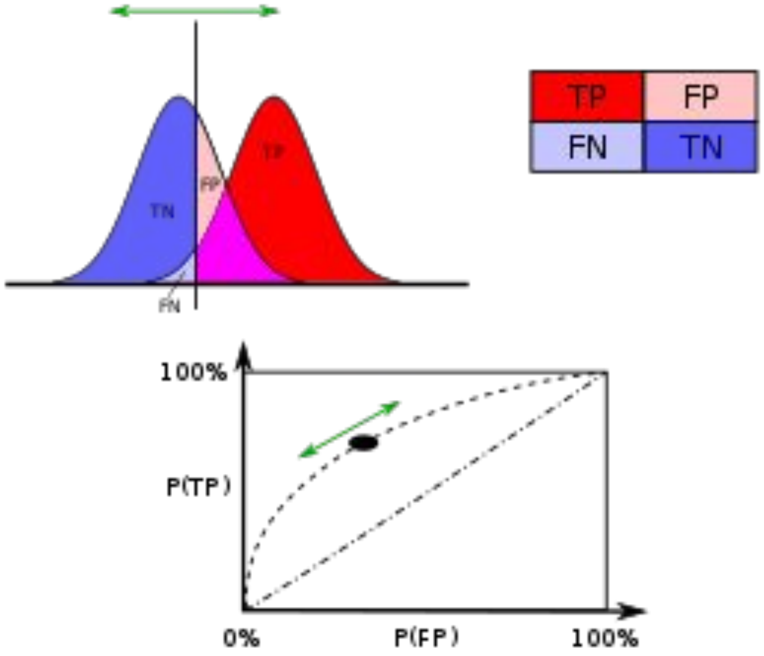
Source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

- Relationship between TPR and FPR
- The diagonal is chance classification (no predictive ability)

- Threshold for P(Y=1|x): 0
- No FN, no TN (all positive predictions)
- TPR = TP/(TP+FN) = TP/(TP+0) = TP/TP = 100%
- FPR = FP/(FP+TN) = FP/(FP+0) = FP/FP = 100%

		True observation	
		1	0
Prediction	1	TP	FP
	0	0 (FN)	0 (TN)

ROC curves



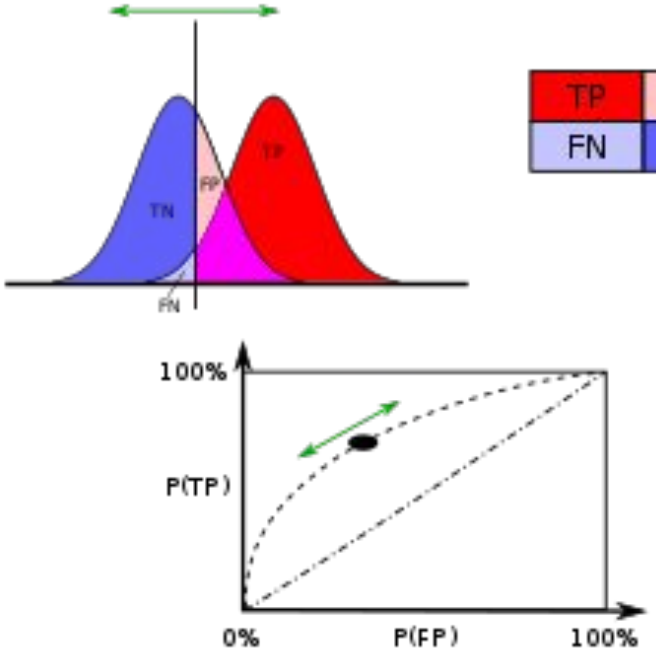
Source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

- Relationship between TPR and FPR
- The diagonal is chance classification (no predictive ability)

- Threshold for P(Y=1|x): 1
- No TP, no FP (all negative predictions)
- TPR = (TP=0)/(TP=0+FN) = 0%;
- FPR = (FP=0)/(FP=0+TN) = 0%

		True observation	
		1	0
Prediction	1	0	0
	0	FN	TN

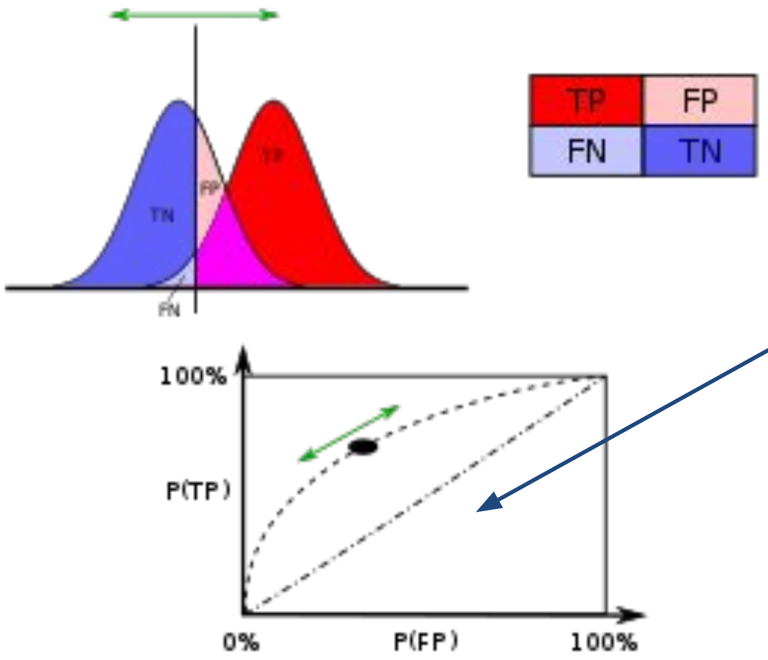
ROC curves



- Relationship between TPR and FPR
- The diagonal is chance classification (no predictive ability)
- **The best is towards the left upper corner (TPR → 100%, FPR → 0%)**

Source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

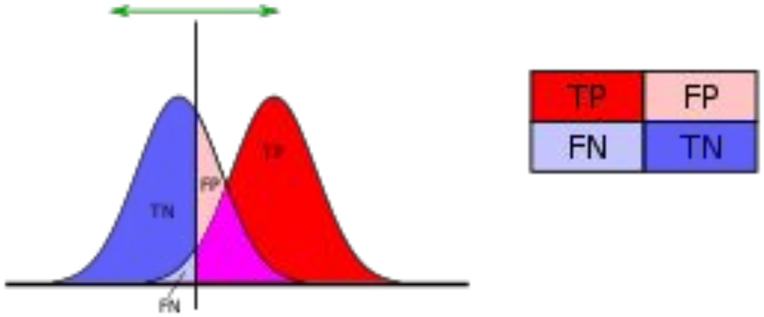
Area under the curve (AUC)

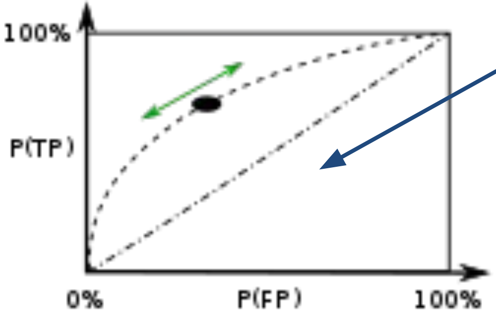


- AUC = 0.5: random guessing
- AUC = 1: perfect classifier
- AUC > 0.8: good classifier

Source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Area under the curve (AUC)



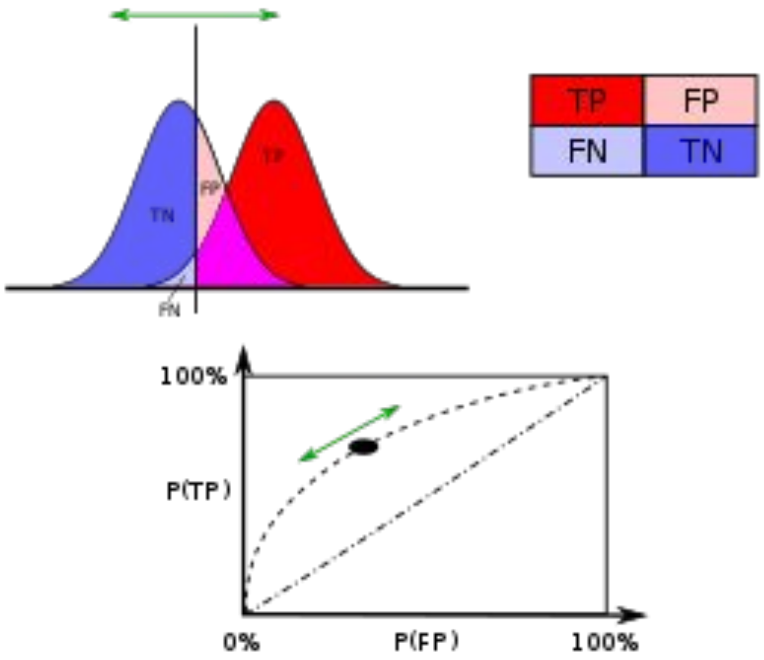


- AUC = 0.5: random guessing
- AUC = 1: perfect classifier
- AUC > 0.8: good classifier

AUC = 0.8 → 80% chance that a true positive sample will have higher probability of being classified as positive than a true negative

Source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Cut-off thresholds



- Two types of error: **FP**, **FN**: sometimes one error may be more critical than the other
- e.g. for a bank may be more important to correctly identify borrowers who will default at the expense of an increase of false positives (and of the total error rate) → lower cut-off for P(y=1|x)
- e.g. in a pandemic, you may want to be sure about detecting carriers, even if this means increasing the FPR

Source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

ROC AUC: limitations

- strongly unbalanced data (tot accuracy = 0.97)
- AUC: looks at TPR and FPR (perspective from actual labels):
    - TPR = 161/(161+6) = 0.96
    - FPR = 0/(0+12) = 0 → (TNR = 1)
- AUC can be close to 1 (depends on distribution of probabilities)
- doubling the n. of false negatives (6 → 12) would change TPR to be 0.93 (still high) (FPR is still 0, AUC can still be close to 1)

	observed labels	
predictions	**neg**	**pos**
neg	12	6
pos	0	161

[by columns]

ROC AUC: limitations

- strongly unbalanced data (tot accuracy = 0.97)
- AUC: looks at TPR and FPR:
    - TPR = 161/(161+6) = 0.96
    - FPR = 0/(0+12) = 0 → (TNR = 1)
- however:
    - PPV = 161/(161+0) = 1 (FDR = 1 - PPV = 0)
    - NPV = 6/(6+12) = 0.667 (FOR = 1- NPV = 0.333)
- it would be nice to have a metric that looks at all four rates: TPR, TNR, PPV, NPV

  →**Matthews Correlation Coefficient**

	observed labels	
predictions	**neg**	**pos**
neg	12	6
pos	0	161

[by rows]

MCC: Matthews Correlation Coefficient

$$\phi = \frac{(TP \cdot TN - FP \cdot FN)}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$$

	observed labels	
predictions	**neg**	**pos**
neg	12	6
pos	0	161

[by rows]

- **range: [-1, +1]**
  - **−1**: total disagreement between predicted classes and actual classes
  - **0**: complete random guessing (no predictive ability)
  - **+1**: total agreement between predicted classes and actual classes

MCC: Matthews Correlation Coefficient

$$\phi = \frac{(TP \cdot TN - FP \cdot FN)}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$$

	observed labels	
predictions	**neg**	**pos**
neg	12	6
pos	0	161

[by rows]

- TP: 161
- TN: 12
- FP: 0
- FN: 6

MCC = (161*12 - 0*6) / sqrt((161+0)*(161+6)*(12+0)*(12+6))

MCC = 1932/2409.895 = 0.802

ROC curves

- demonstration 4.2

→ 4.classification.Rmd