

Supervised learning: classification problems

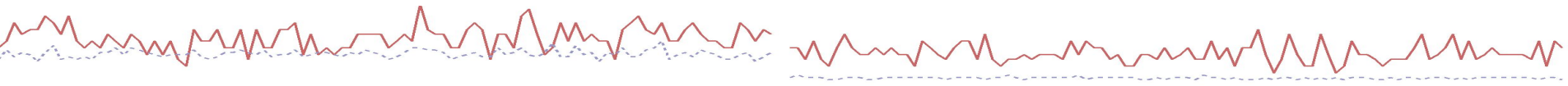
Filippo Biscarini (CNR, Milan, Italy)

filippo.biscarini@cnr.it



Classification problems

- the response variable **y** is **qualitative**
- e.g.: coat colour, type of rice (Tropical japonica, Indica, Temperate japonica, Aromatic, Aus)
- special case → binary traits (e.g. cases/controls, resistant/susceptible)
- **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:

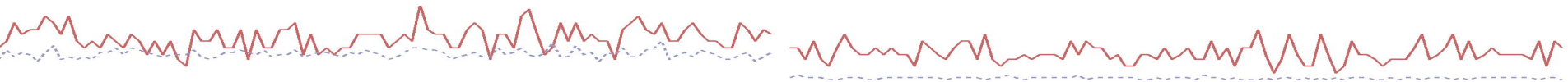
- K classes ($k \in K$)

probabilities

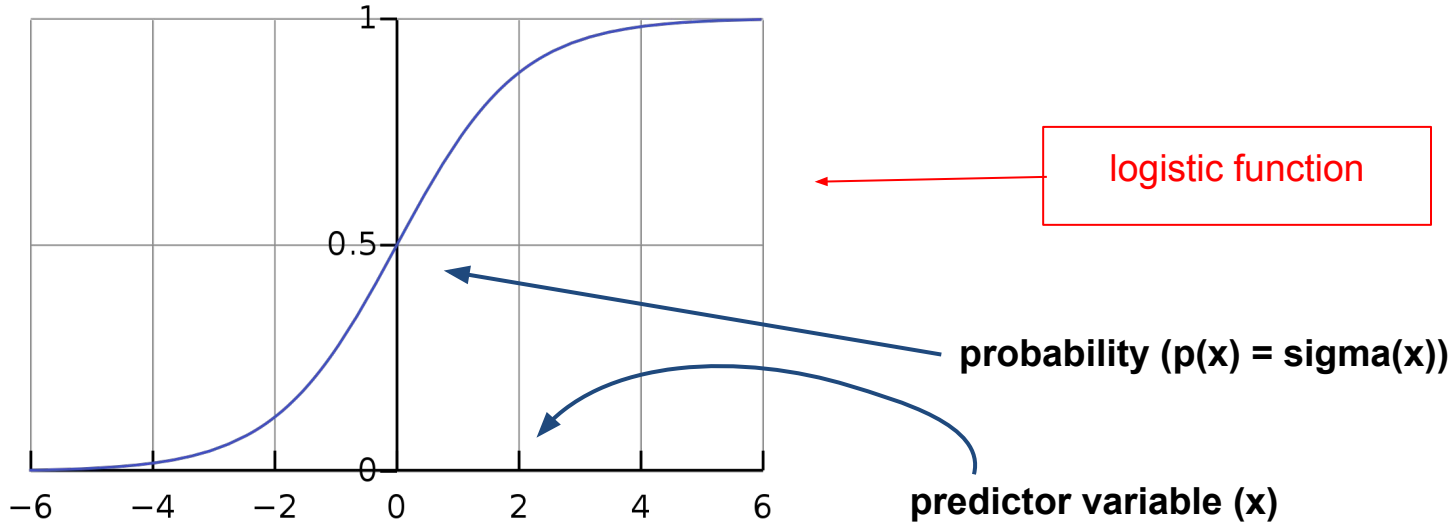
classifier

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

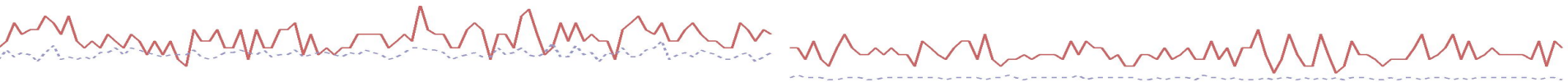
$$C(x) = k, \text{ if } p_k(x) = \max\{p_1(x), p_2(x), \dots, p_K(x)\}$$



Binary classification problems



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

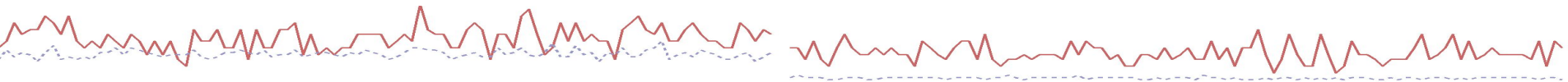


Logistic regression

- the logistic function is the basis for **logistic regression**
- $P(y=1|x)$ [also $p(x)$]
- $Z = \beta_0 + \beta_1 x$ (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

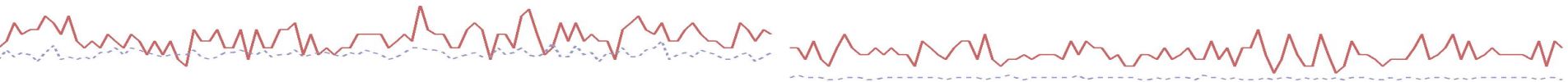


Logistic regression

- 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

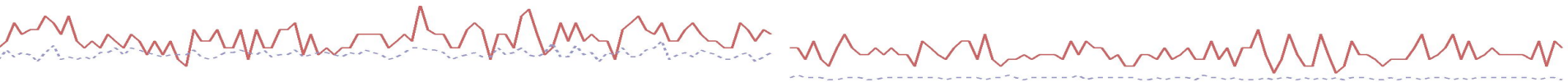
- a little bit of algebra:

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

odds

log(odds): logit

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



Logistic regression

- the **logit function** ($\log(\text{odds})$) is the **link function** between a linear expression of X and the probabilities of Y
- linear X expression $(\beta_0 + \beta_1 x) \rightarrow$ logit scale (continuous)
- logistic function: converts values on the logit scale back to probabilities

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

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
\mathbf{y}	$\hat{y} = \sigma(\beta_0 + \beta_1 x)$

difference between observed and
predicted values



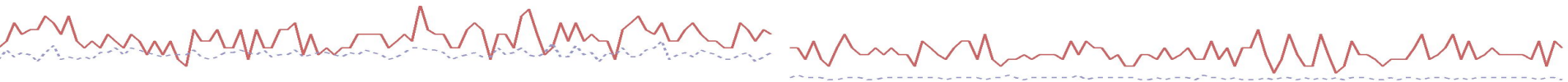
Estimating the coefficients

how do we obtain the model coefficients β ?

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

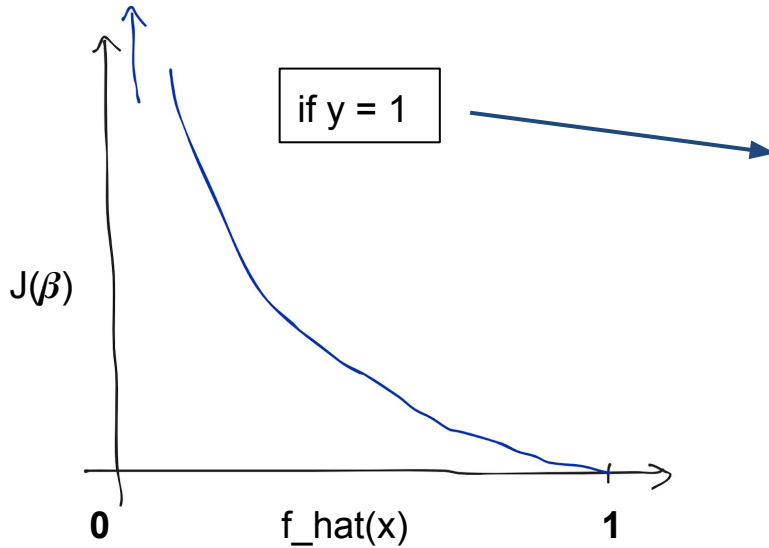
$$J(\beta) = \text{Cost}(\hat{y}, y) = \begin{cases} -\log(\hat{y}) & \text{if } y = 1 \\ -\log(1 - \hat{y}) & \text{if } y = 0 \end{cases}$$

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



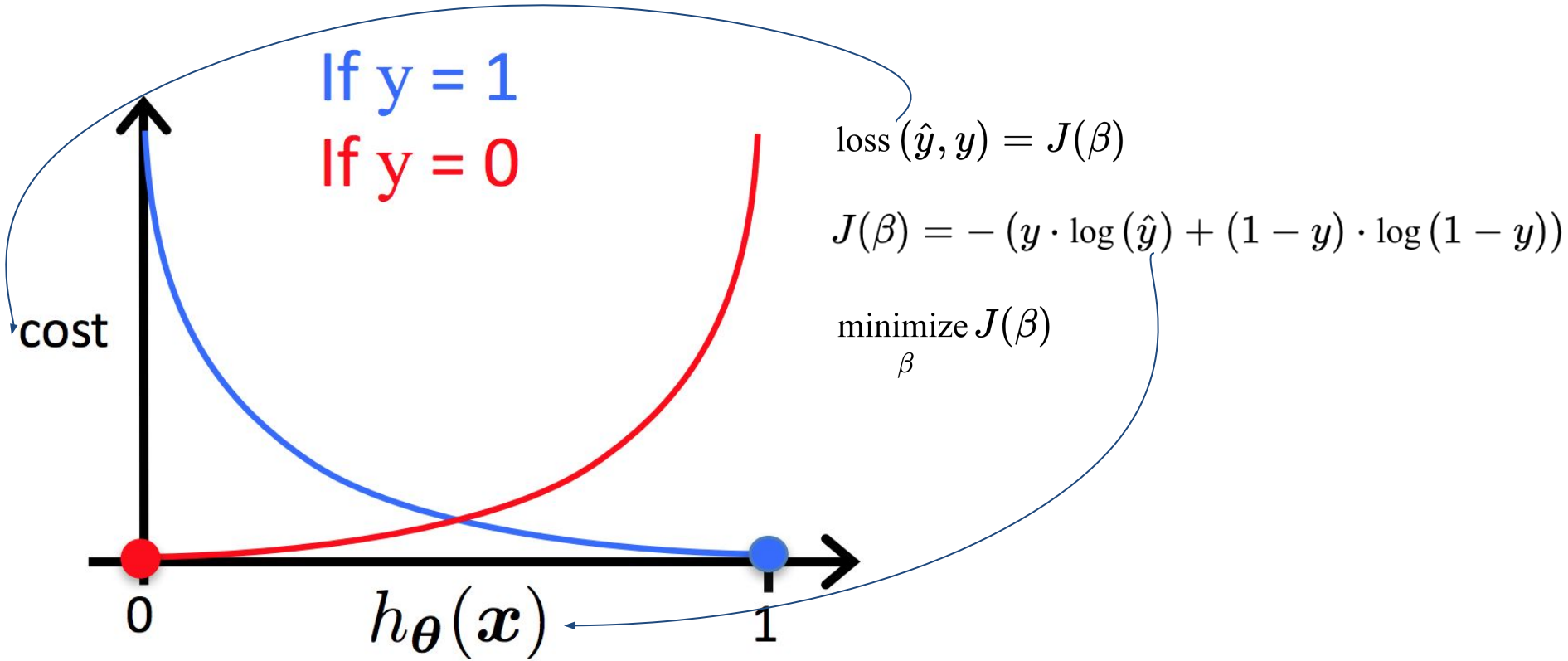
Cost function for logistic regression

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



- if $y_{\text{hat}} = 1$, cost = 0
- if $y_{\text{hat}} \rightarrow 0$ (but $y = 1$), cost \rightarrow infinity
- the opposite holds if $y = 0$

Loss function for logistic regression



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: measuring performance

- 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

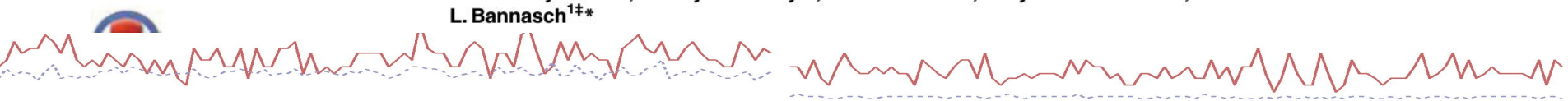
binary phenotypes: **cleft lip** (presence/absence)



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☯na}, John R. Shaffer^{3☯}, Elizabeth J. Leslie², Boaz Arzi⁴, Cali E. Willet⁵, Timothy C. Cox^{6,7,8}, Toby McHenry², Nicole Narayan⁹, Eleanor Feingold³, Xioajing Wang^{2na}, Saundra Sliskovic¹, Nili Karmi¹, Noa Safra¹, Carla Sanchez², Frederic W. B. Deleyiannis¹⁰, Jeffrey C. Murray¹¹, Claire M. Wade⁵, Mary L. Marazita^{2,12‡*}, Danika L. Bannasch^{1‡*}



Genetic variants for cleft lip in dogs

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

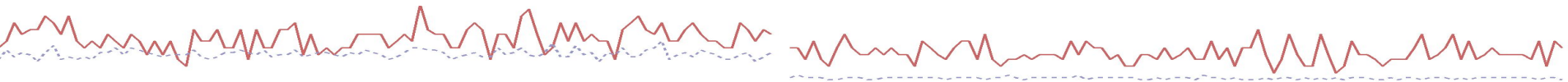
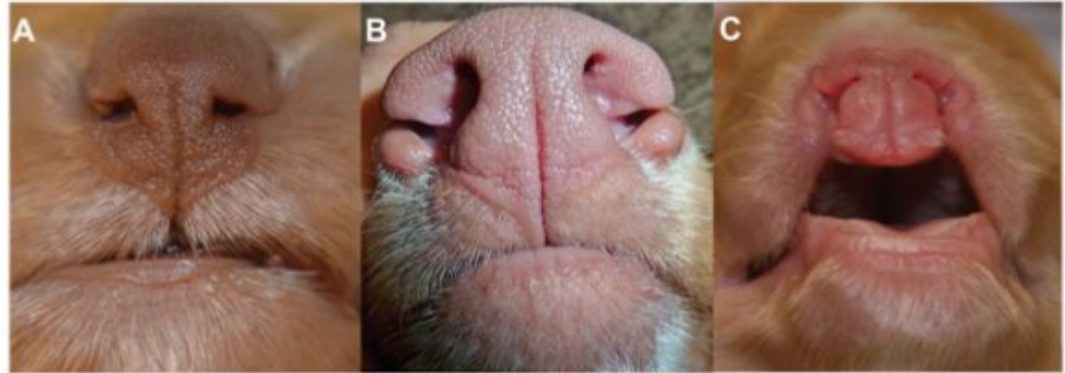
- Nova Scotia Duck Tolling Retriever (NSDTR)
- 125 dogs:
 - 13 cases
 - 112 controls



Genetic variants for cleft lip in dogs

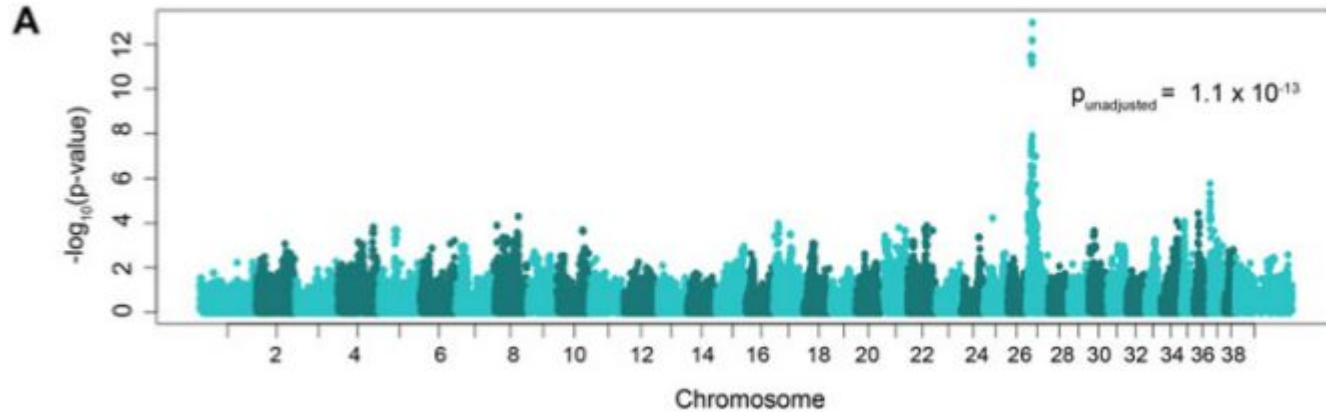
binary phenotypes: **cleft lip** (presence/absence)

- 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

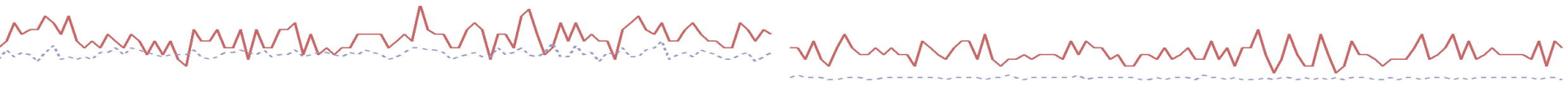
Strong signal of
association on
chromosome 27



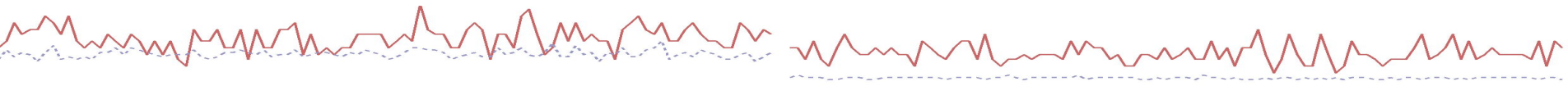
Logistic regression

- demonstration 4.1
- Exercise 4.1

→ `4.classification.ipynb`



ROC curves



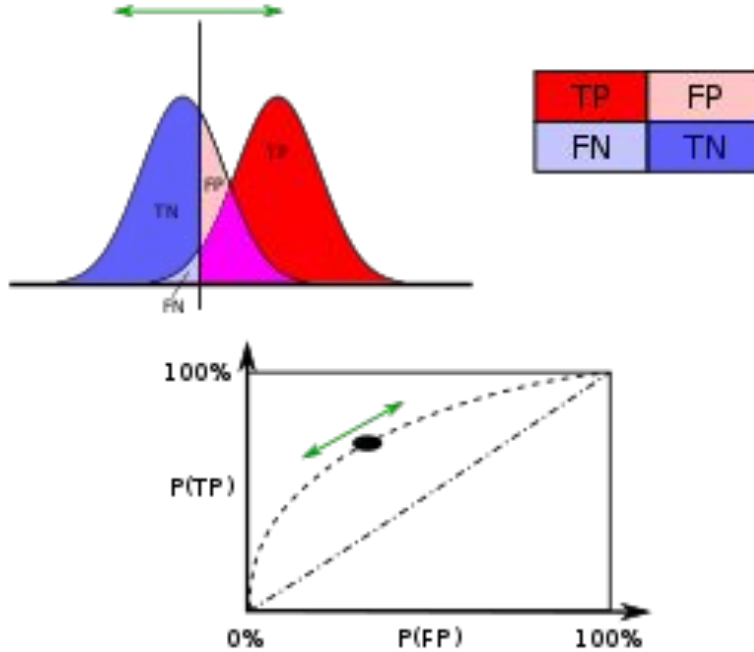
Binary classification

		True observation	
		1	0
Prediction	1	TP	FP
	0	FN	TN

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

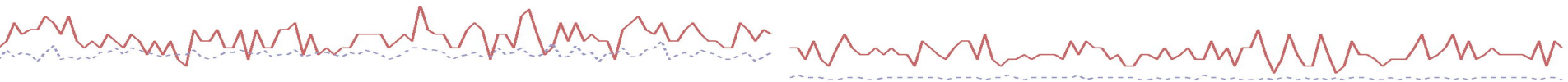


ROC curves

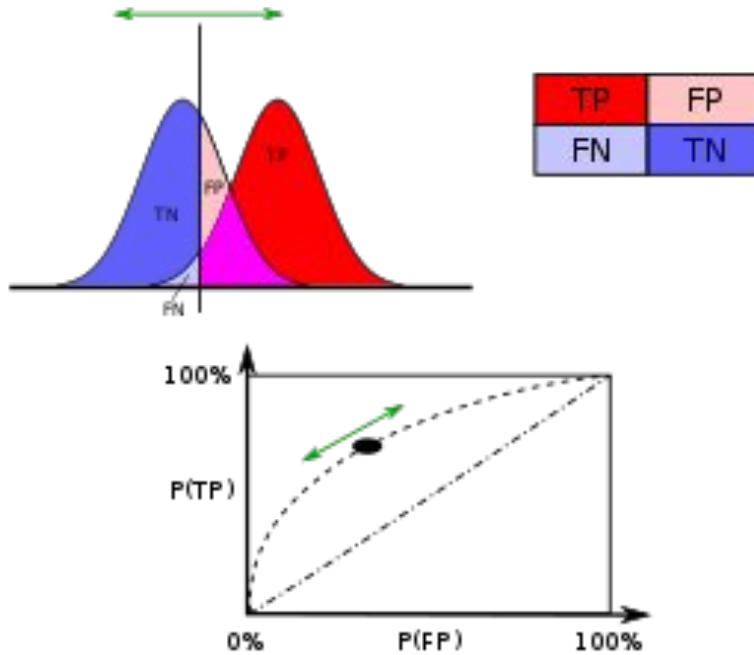


- Relationship between TPR and FPR
- The diagonal is chance classification (no predictive ability)

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



ROC curves



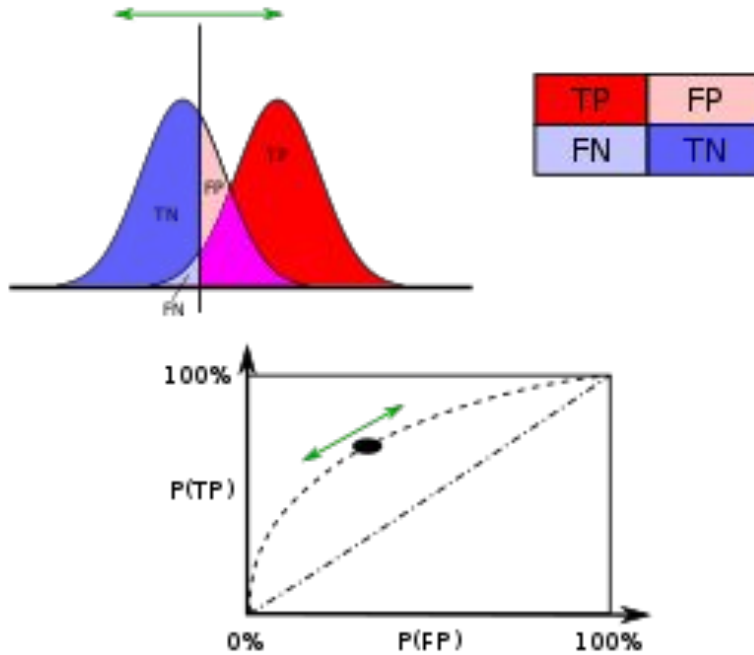
- Relationship between TPR and FPR
- The diagonal is chance classification (no predictive ability)

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

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

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

ROC curves



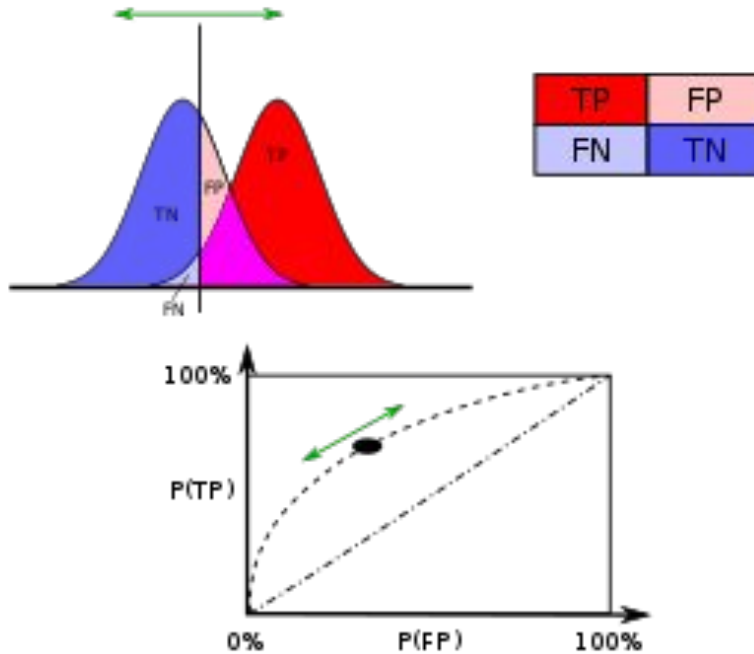
- Relationship between TPR and FPR
- The diagonal is chance classification (no predictive ability)

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

		True observation	
		1	0
Prediction	1	0	0
	0	FN	TN

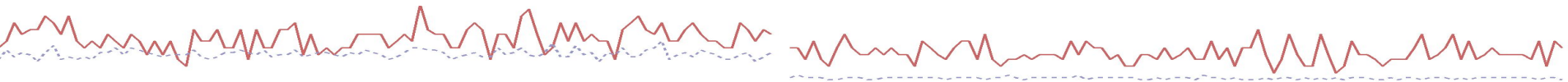
Source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

ROC curves

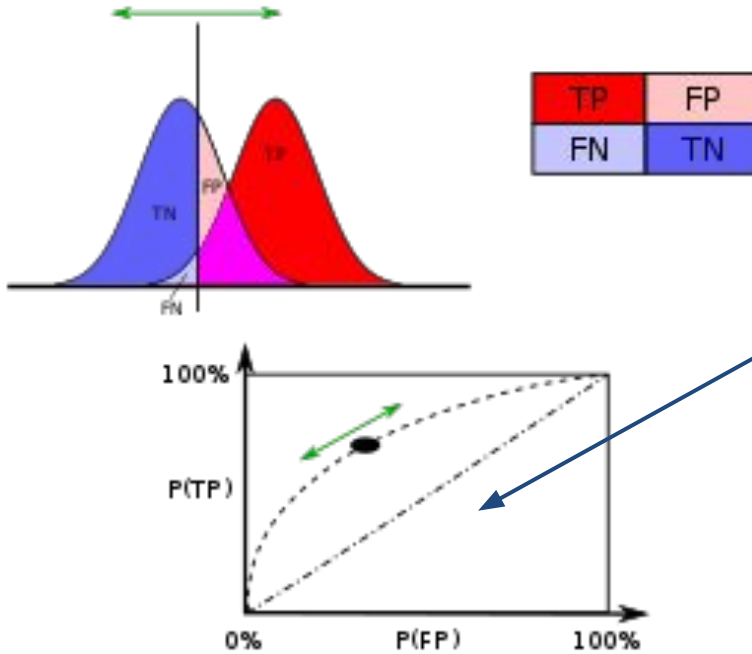


- Relationship between TPR and FPR
- The diagonal is chance classification (no predictive ability)
- **The best is towards the left upper corner (TPR \rightarrow 100%, FPR \rightarrow 0%)**

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



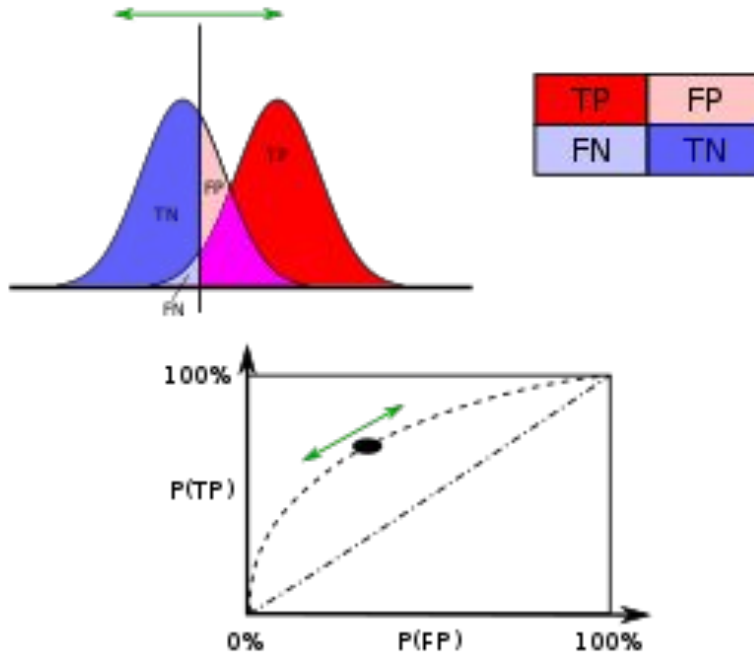
Area under the curve (AUC)



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

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

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 curves

- demonstration 4.2

→ 4.classification.ipynb

