Разбор домашнего задания №1 по теме реализации kNN

Young&&Yandex ШАД

girafe ai

Arkadii Lysiakov

Outline

01 Recap k Nearest Neighbours (kNN)

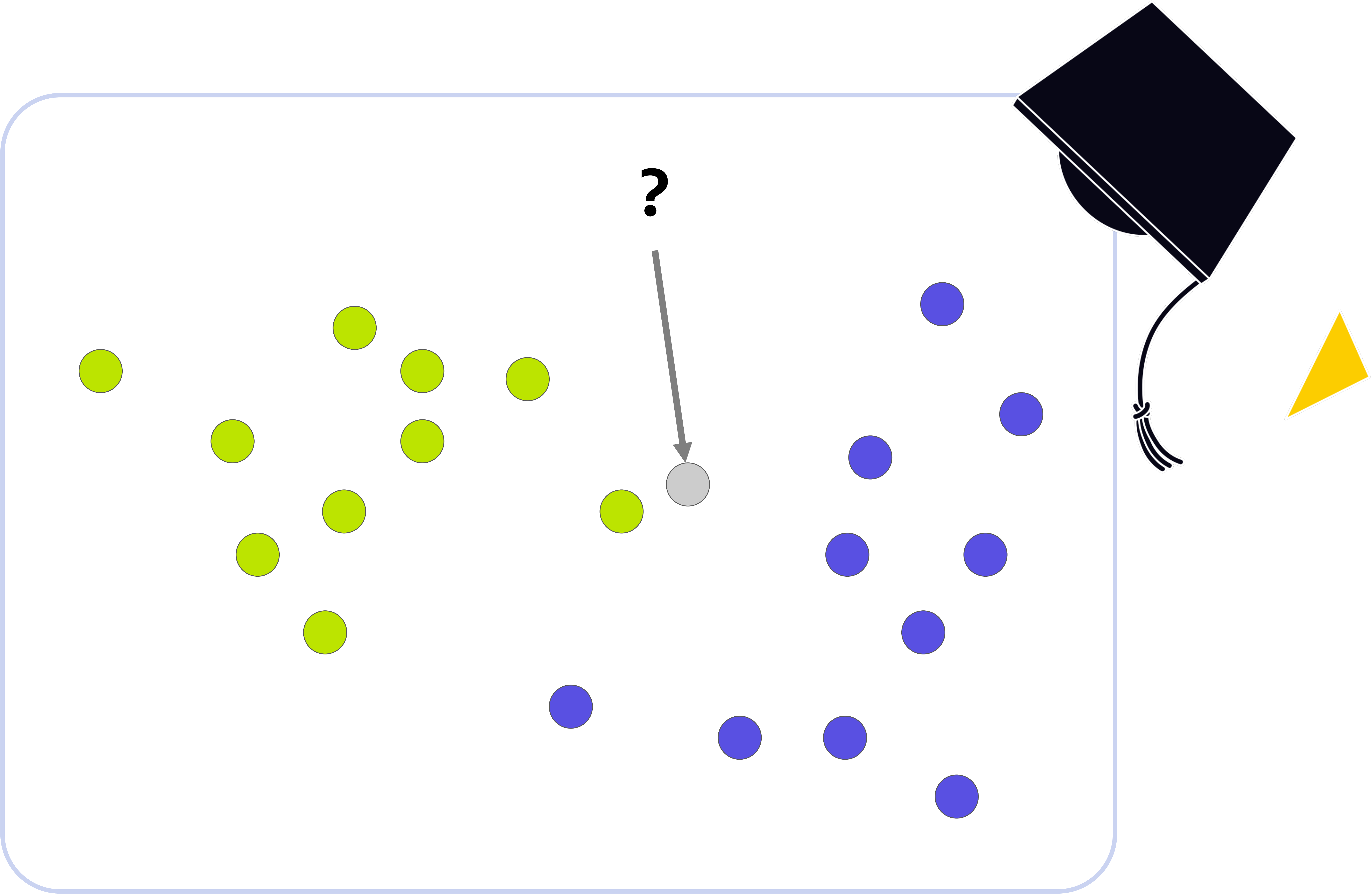
02 Lifecode

03 Recap Maximum Likelihood Estimation:

- Classification
- Regression

04 Bonus part: GLM

kNN — k Nearest Neighbours



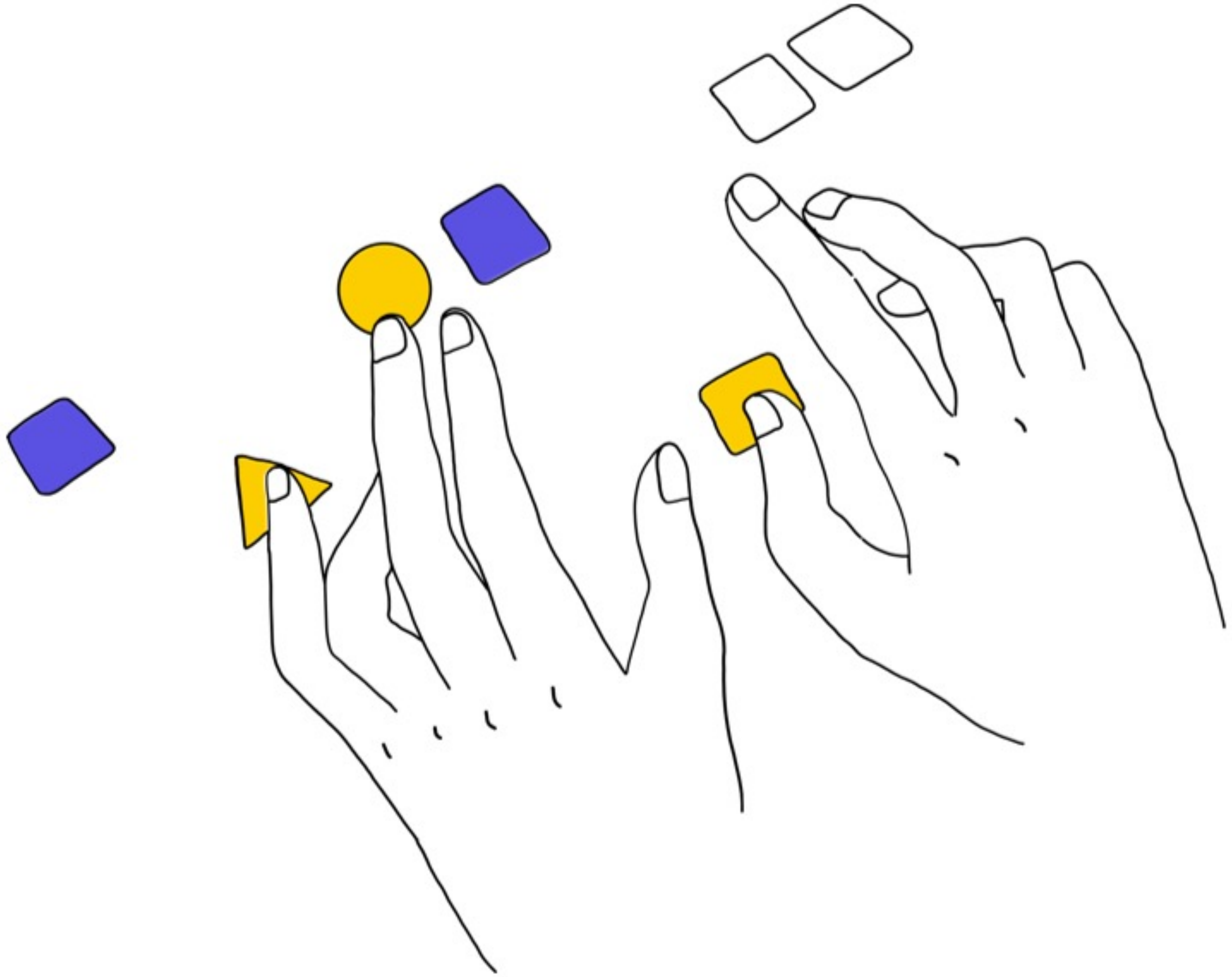
k Nearest Neighbors

Given a new observation:

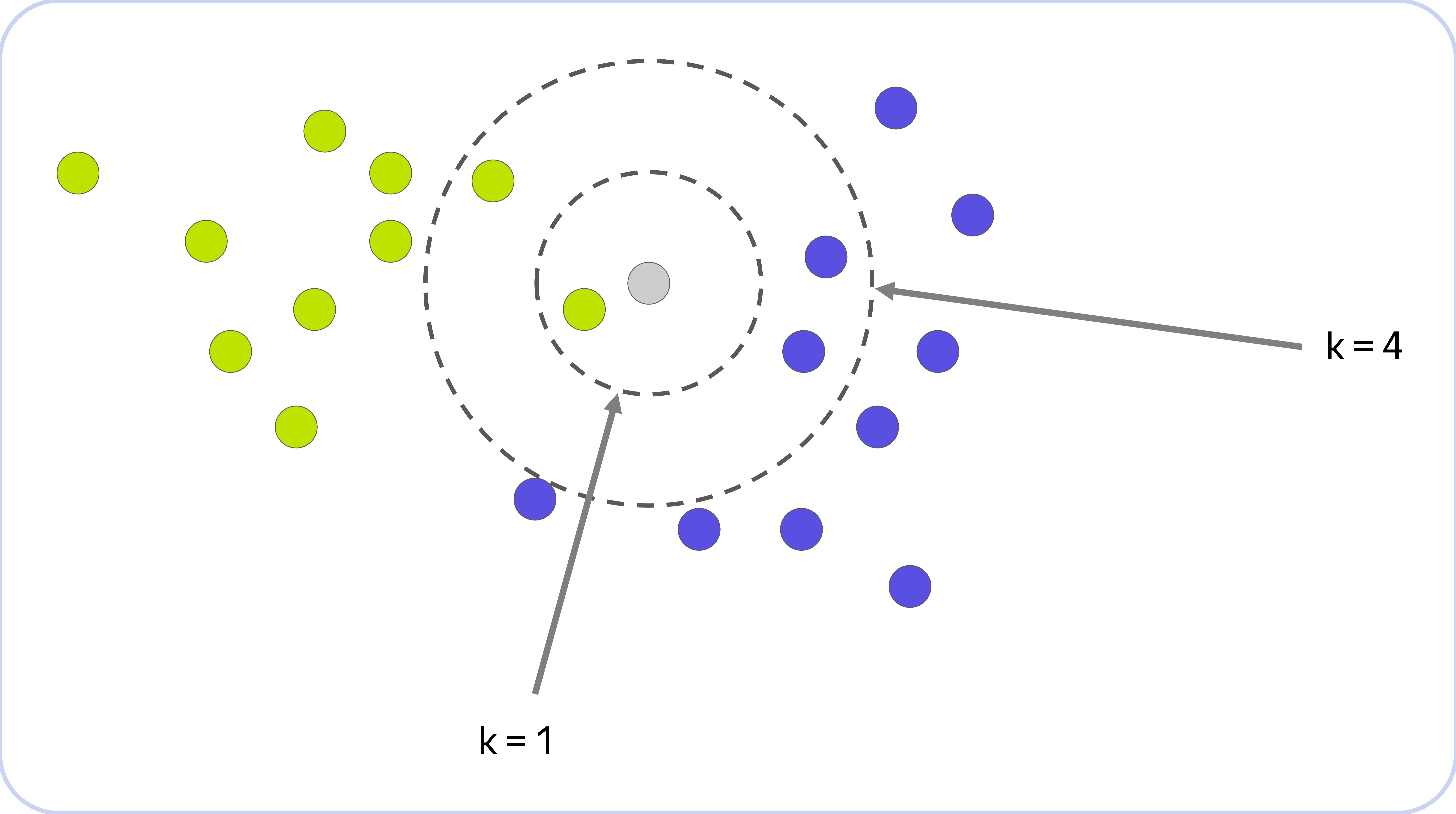
01 Calculate the distance to each of the samples in the dataset

02 Select samples from the dataset with the minimal distance to them

03 The label of the new observation will be the most frequent label among those nearest neighbors



kNN — k Nearest Neighbours



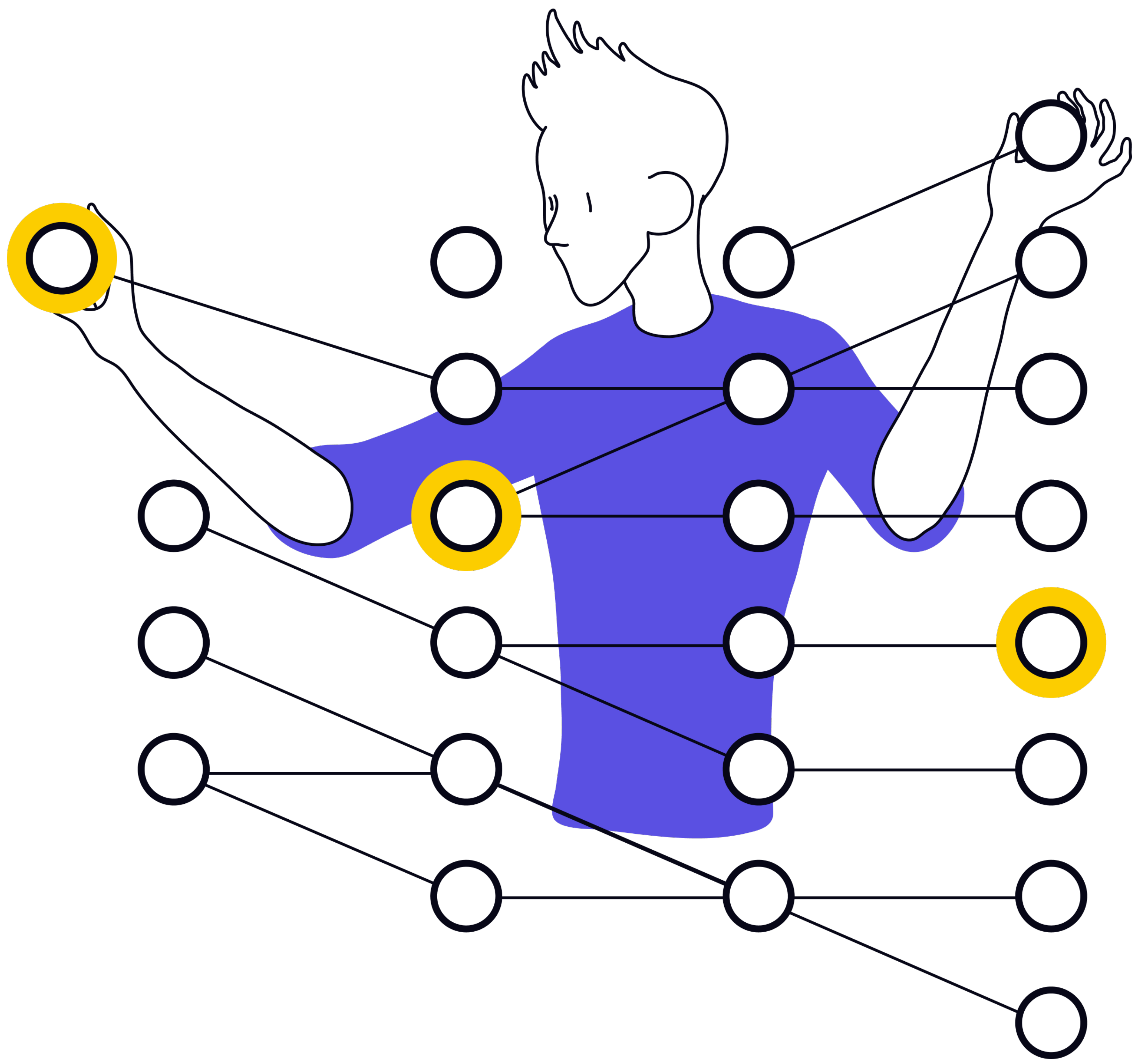
How to make it better?

01 The number of neighbors k
(it is a **hyperparameter**)

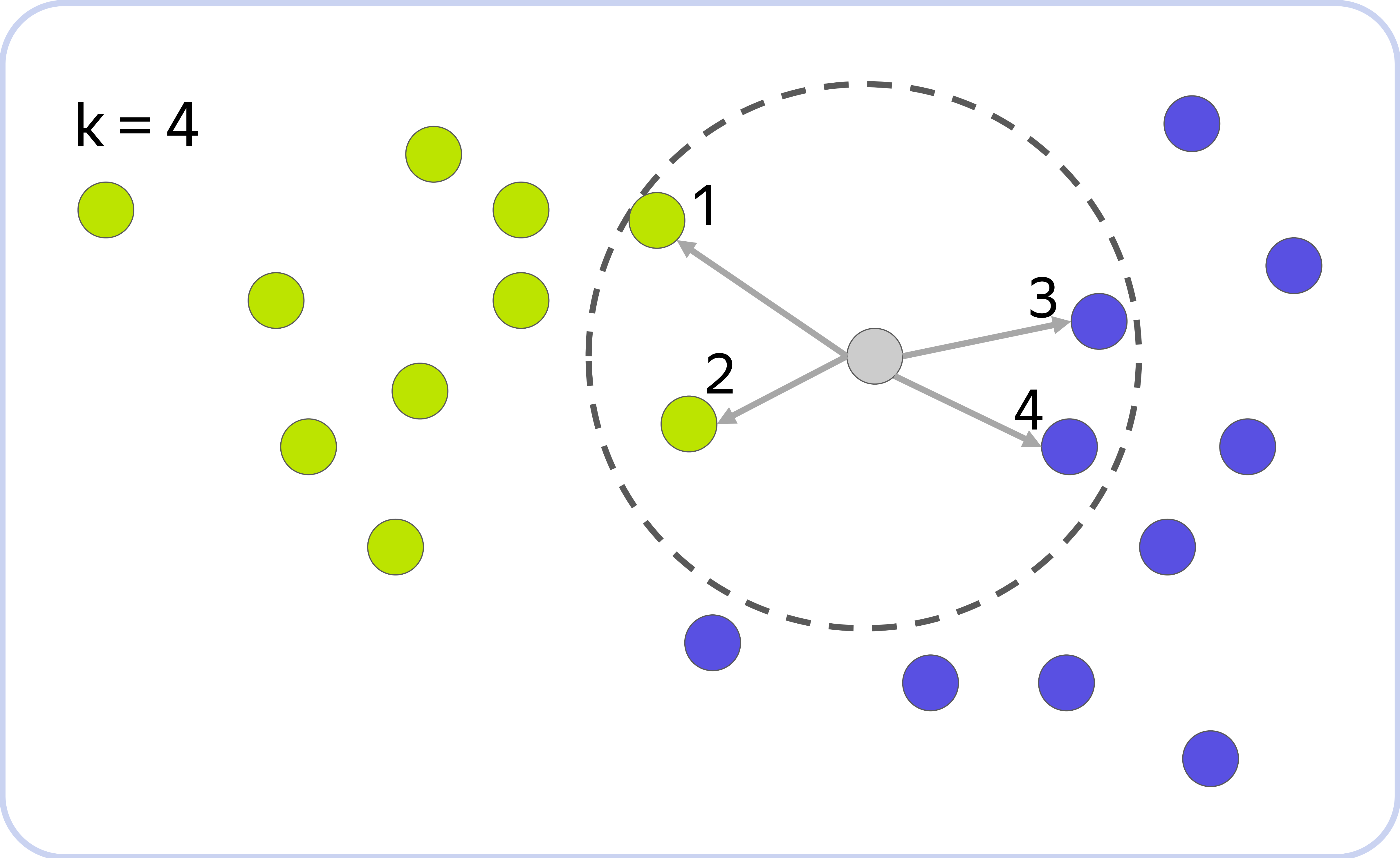
02 The distance measure between samples

- Hamming
- Euclidean
- Cosine
- Minkowski distances
- etc.

03 Weighted
neighbours



Weighted kNN



Weights can be adjusted according to the neighbors order,

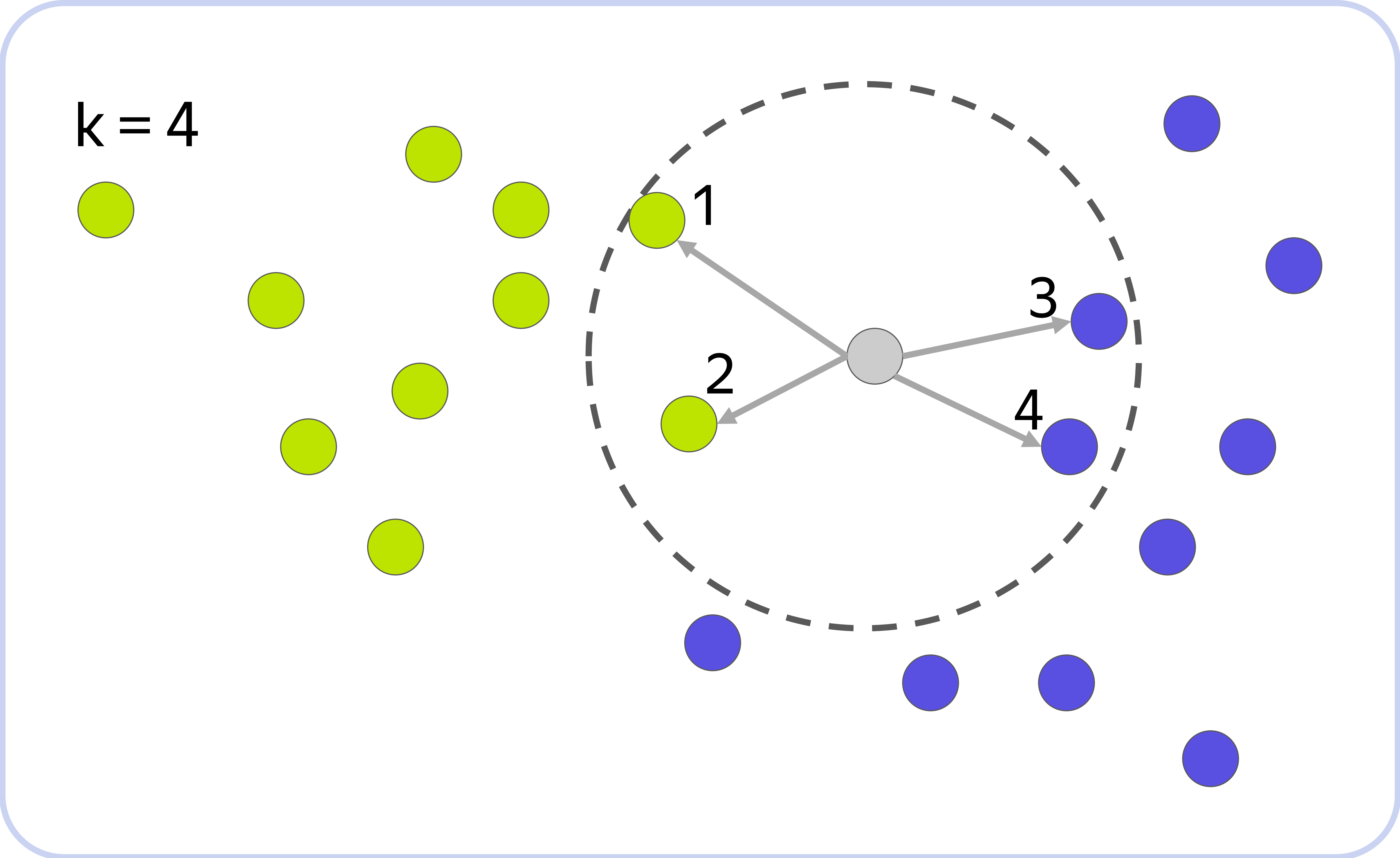
$$w(\boldsymbol{x}^{(i)}) = w\big(d(\boldsymbol{x}^{(i)}, \boldsymbol{x})\big)$$

or on the distance itself

$$w(\boldsymbol{x}^{(i)}) = w_i$$

$$p_{\text{green}} = \frac{w(\boldsymbol{x}^{(1)}) + w(\boldsymbol{x}^{(2)})}{w(\boldsymbol{x}^{(1)}) + w(\boldsymbol{x}^{(2)} + w(\boldsymbol{x}^{(3)}) + w(\boldsymbol{x}^{(4)})}$$

Weighted kNN



Weights can be adjusted according to the neighbors order,

$$w(\boldsymbol{x}^{(i)}) = w_i$$

or on the distance itself

$$w(\boldsymbol{x}^{(i)}) = w\big(d(\boldsymbol{x}^{(i)}, \boldsymbol{x})\big)$$

$$p_{\text{blue}} = \frac{w(\boldsymbol{x}^{(3)}) + w(\boldsymbol{x}^{(4)})}{w(\boldsymbol{x}^{(1)}) + w(\boldsymbol{x}^{(2)} + w(\boldsymbol{x}^{(3)}) + w(\boldsymbol{x}^{(4)})}$$

Lifecode

Likelihood

Denote dataset generated by distribution with parameter θ

Likelihood function:

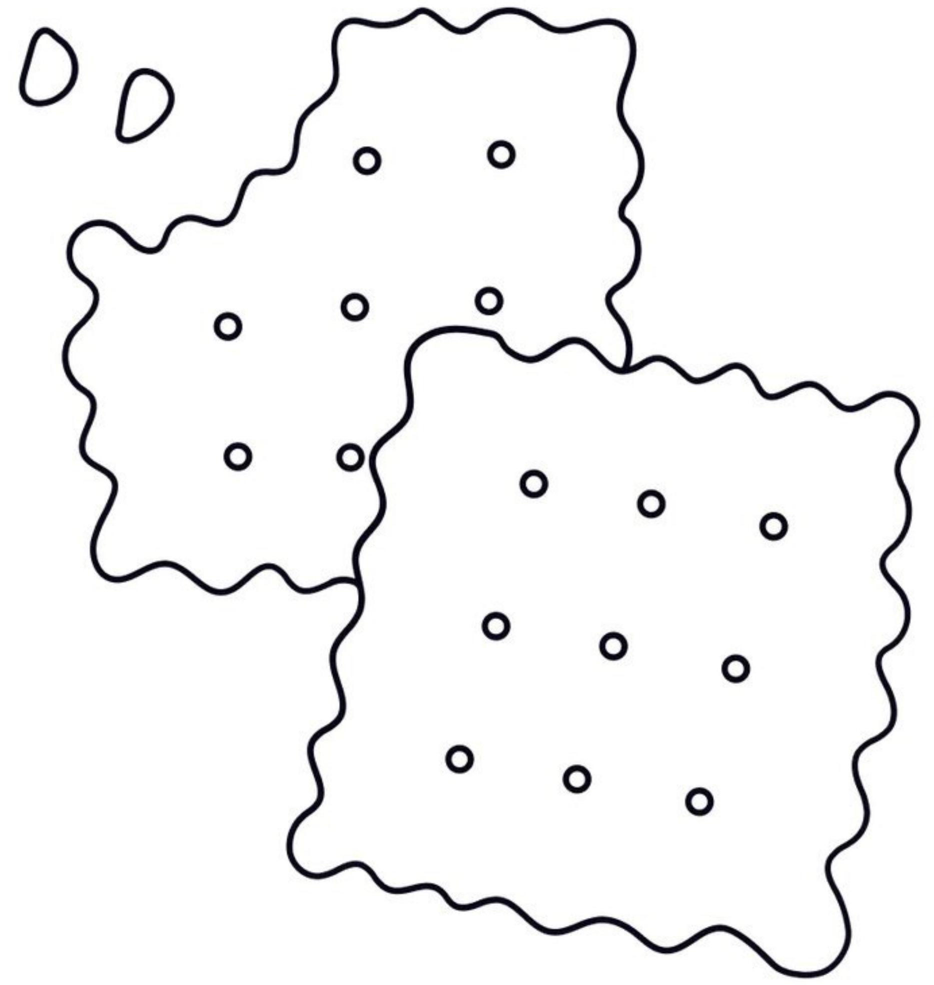
$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y}) = P(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} P(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}|\boldsymbol{\theta})$$

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y}) \longrightarrow \max_{\boldsymbol{\theta}}$$

samples should be i.i.d.

equivalent to

$$\log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{n} \log P(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}|\boldsymbol{\theta}) \longrightarrow \max_{\boldsymbol{\theta}}$$



Classification problem

$$X \in R^{n \times p}$$

$$Y \in C^n \qquad \text{e.g. } C = \{-1, 1\}$$

$$|C| < +\infty$$

$$c(X) = \hat{Y} \approx Y$$

Maximum Likelihood Estimation

Just to remind

$$\log L(w|X,Y) = \log P(X,Y|w) = \log \prod_{i=1}^{n} P(x_i, y_i|w)$$

Calculating probabilities for objects

$$\text{if } y_i = 1: \qquad P(x_i, 1|w) = \sigma_w(x_i) = \sigma_w(M_i)$$

$$\text{if } y_i = -1: \quad P(x_i, -1|w) = 1 - \sigma_w(x_i) = \sigma_w(-x_i) = \sigma_w(M_i)$$

$$\log L(w|X,Y) = \sum_{i=1}^{n} \log \sigma_w(M_i) \boxed{= -\sum_{i=1}^{n} \log(1 + \exp(-M_i)) \to \max_w}$$

Linear regression

01 Dataset $\{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$, where $\boldsymbol{x}^{(i)} \in \mathbb{R}^p, y^{(i)} \in \mathbb{R}$

02 The model is linear:

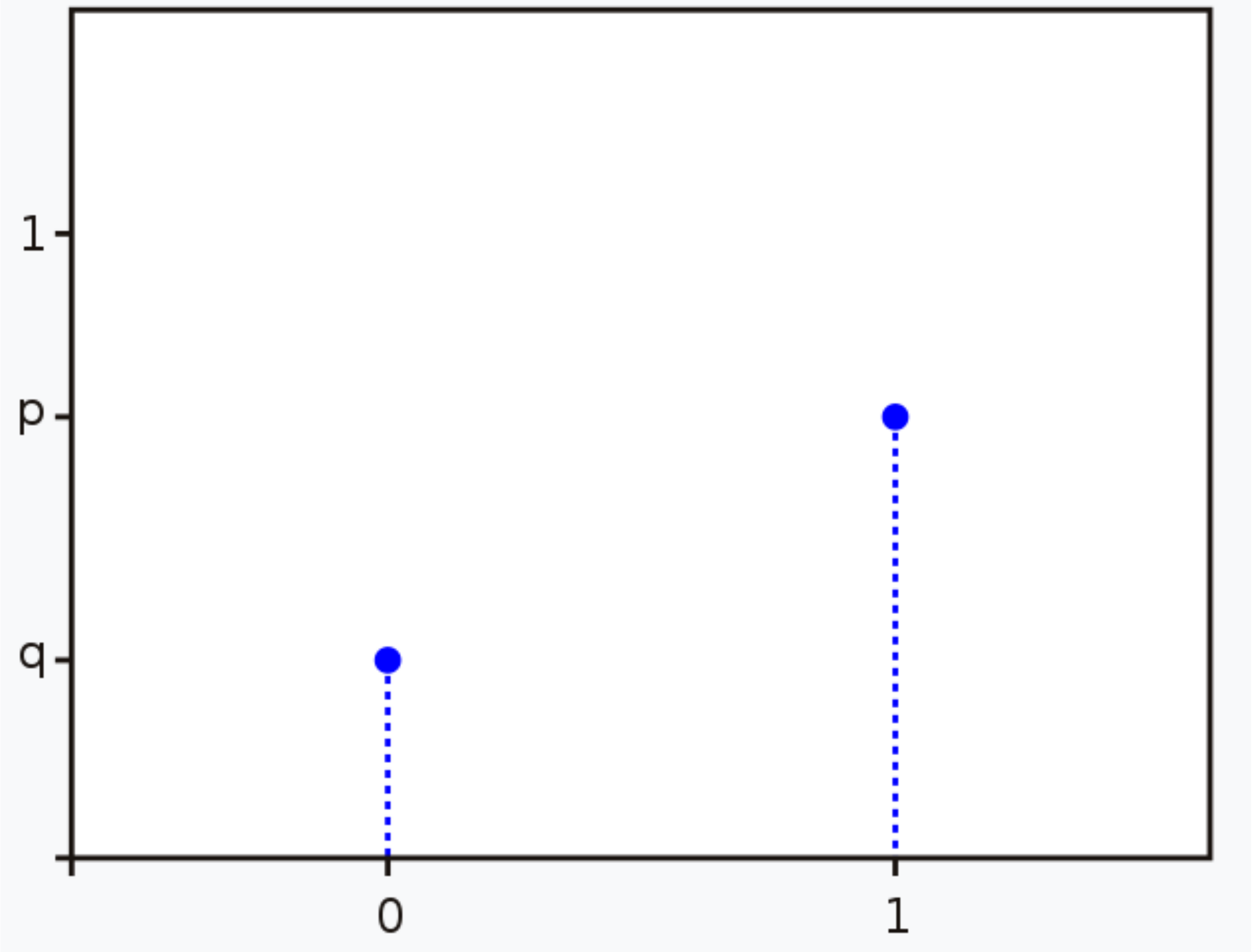
$$\hat{y} = w_0 + \sum_{k=1}^{p} x_k w_k = //\boldsymbol{x} = [1; x_1, \dots, x_p]// = \boldsymbol{x}^\top \boldsymbol{w}$$

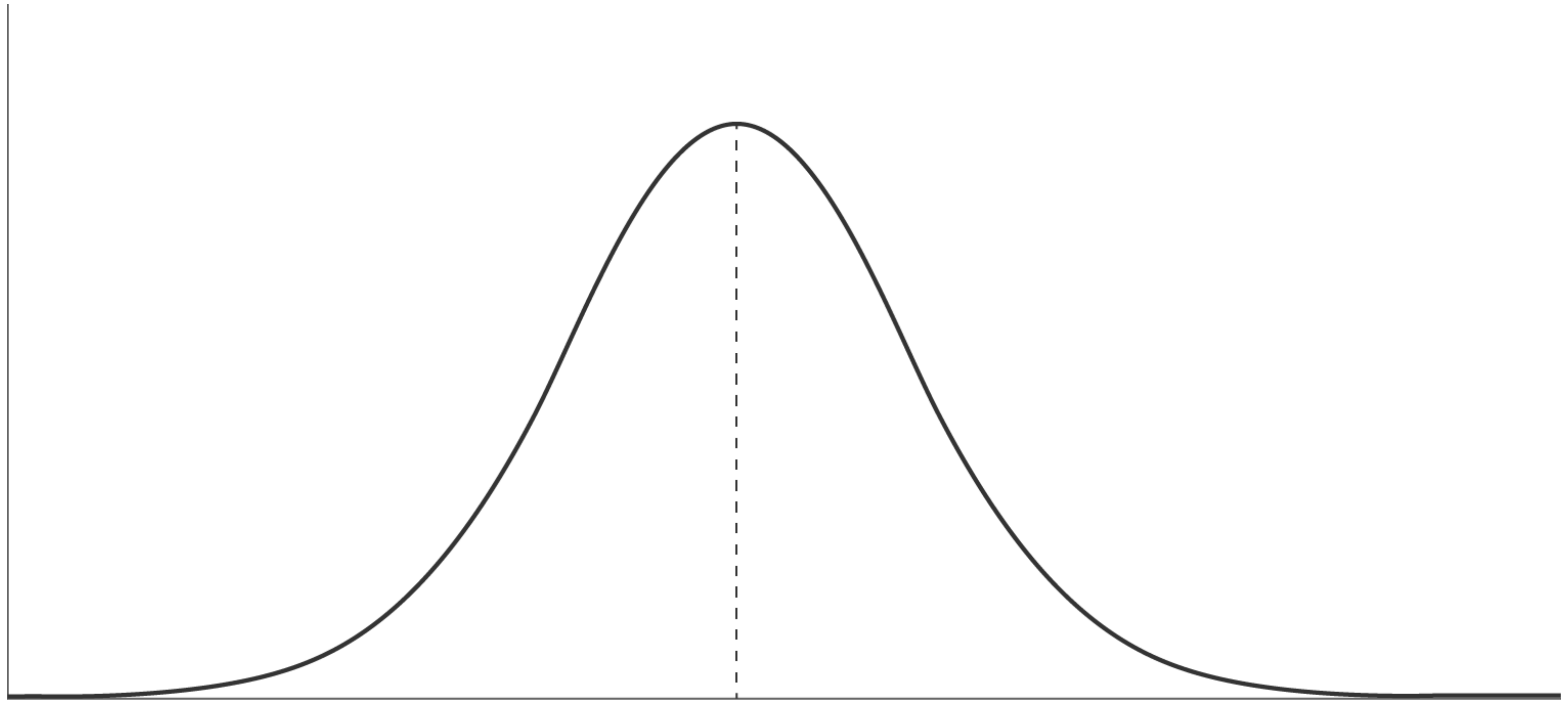
where $\boldsymbol{w} = [w_0; w_1, \dots, w_p]$ is bias term.

03 Least squares method (MSE minimization) provides a solution:

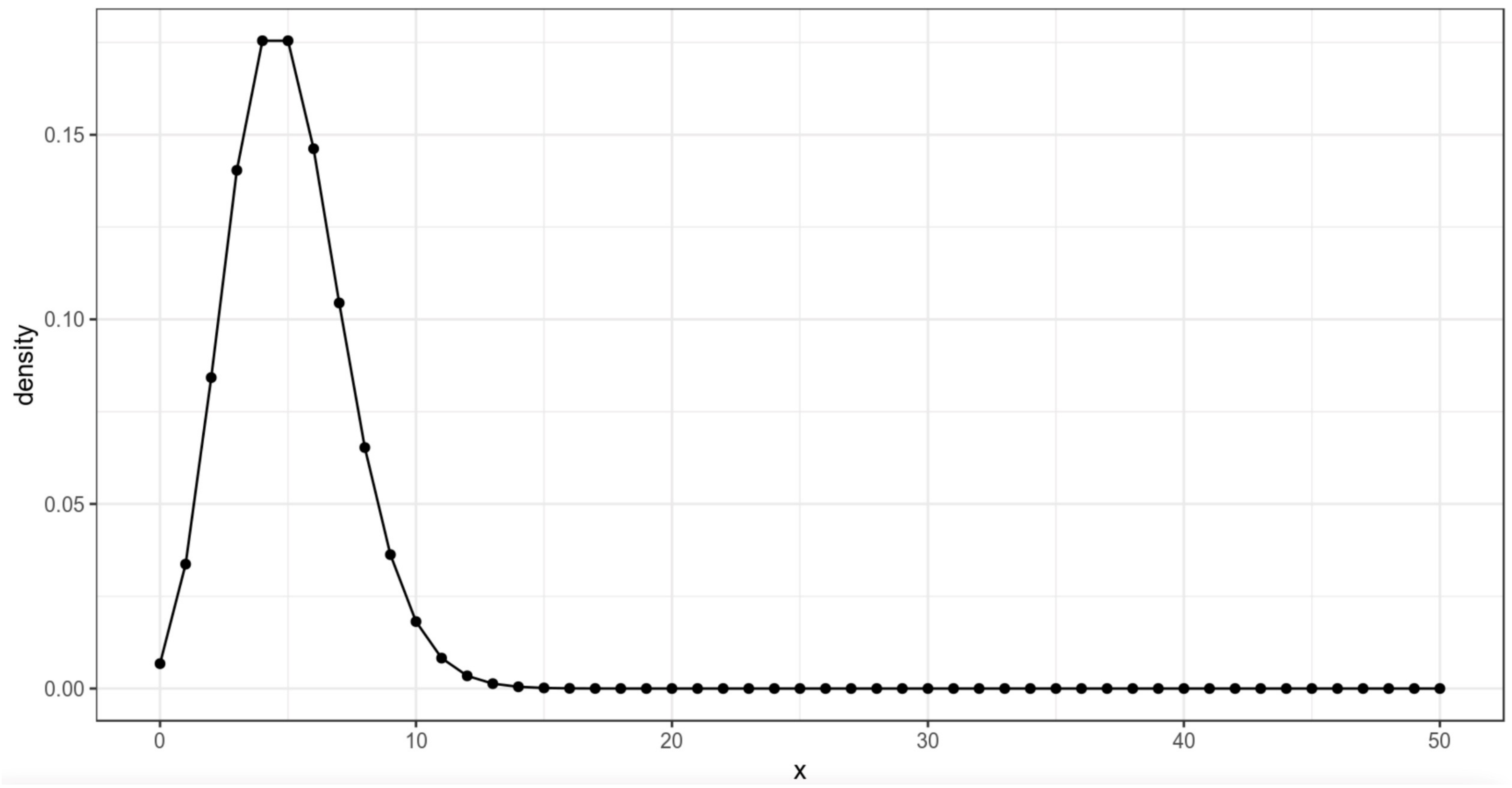
$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{w}||_2^2$$

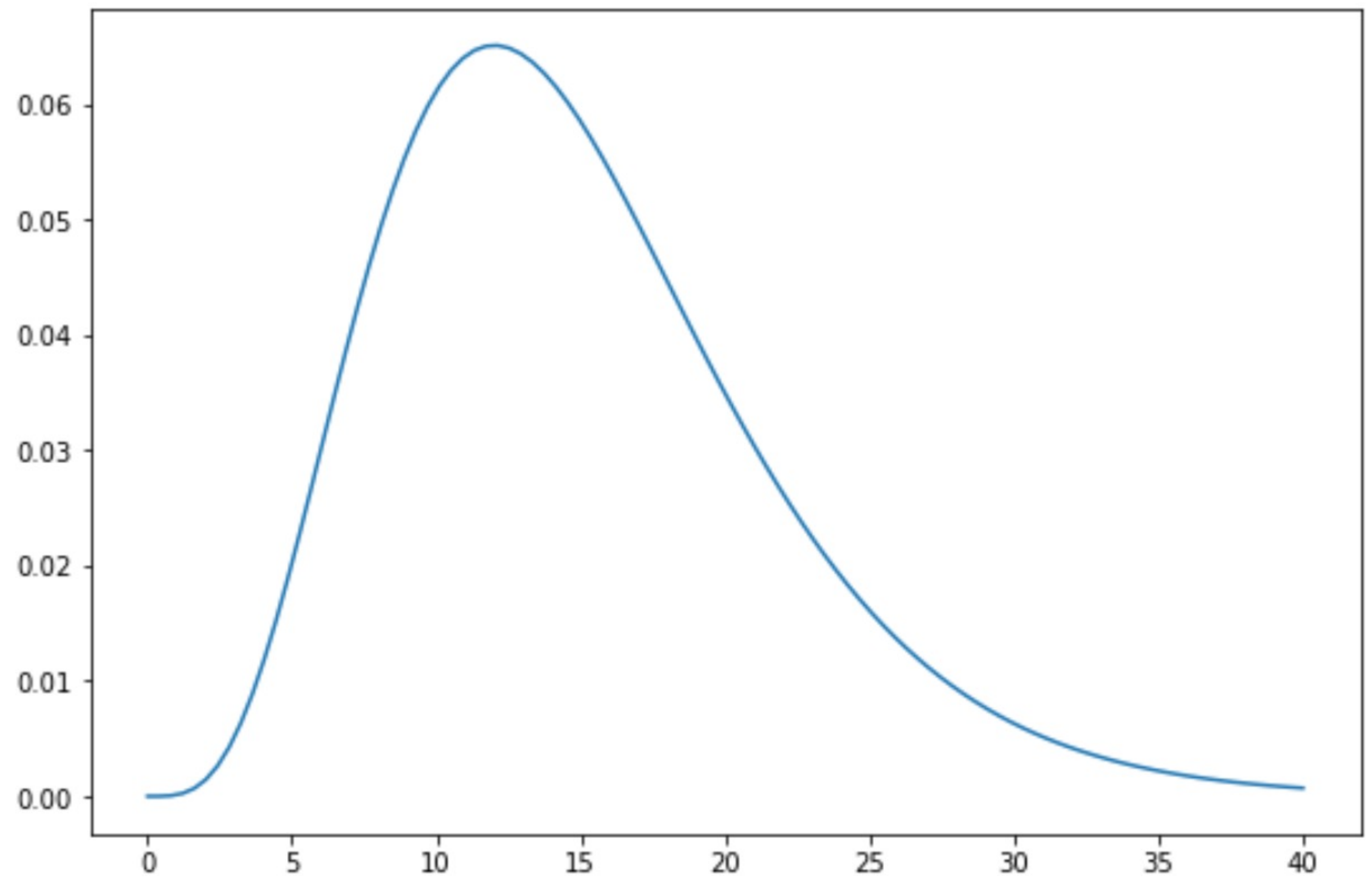
Two distributions





And two other examples





Exponential family

- Normal
- Bernoulli
- Poisson
- Inverse Wishart
- Exponential
- Chi-squared
- Beta
- Wishart
- Gamma
- Dirichlet
- Categorical
- Geometric

Exponential family

Normal	Bernoulli	Poisson
Inverse Wishart	Exponential	Chi-squared
Beta	Wishart	Gamma
Dirichlet	Categorical	Geometric

$$f_X(x \mid \theta) = h(x)\,\exp\left[\,\eta(\theta) \cdot T(x) - A(\theta)\,\right]$$

GLM

01 The GLM consists of three elements

02 A particular distribution for modeling Y from among those which are considered exponential families of probability distributions

03 A linear predictor

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

04 A link function g such that

$$\mathbf{E}(\boldsymbol{Y}\,|\boldsymbol{X}) = \boldsymbol{\mu} = \boldsymbol{g}^{-1}(\boldsymbol{\eta})$$

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function	Mean function
Normal	real: $(-\infty;+\infty)$	Linear-response data	Identity	$X\beta = \mu$	$\mu = X\beta$
Exponential	real: $(0;+\infty)$	Exponential-response data, scale parameters	Negative inverse	$X\beta = -\mu^{-1}$	$\mu = -(X\beta)^{-1}$
Gamma					
Inverse Gaussian	real: $(0;+\infty)$		Inverse squared	$X\beta = \mu^{-2}$	$\mu = (X\beta)^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	Count of occurrences in fixed amount of time/space	Log	$X\beta = \ln(\mu)$	$\mu = \exp(X\beta)$
Bernoulli	integer: $\{0, 1\}$	Outcome of single yes/no occurrence	Logit	$X\beta = \ln(\frac{\mu}{1-\mu})$	$\mu = \frac{\exp(X\beta)}{1+\exp(X\beta)} + \frac{1}{1+\exp(-X\beta)}$
Binomial	integer: $0, 1, \ldots, N$	Count of # of "yes" occurrences out of N yes/no occurrences		$X\beta = \ln(\frac{\mu}{n-\mu})$	
Categorical	integer: $[0, K)$	Outcome of single K-way occurrence		$X\beta = \ln(\frac{\mu}{1-\mu})$	
Multinomial	K-vector of integer: $[0, K]$	Count of occurrences of different types (1, ..., K) out of N total K-way occurrences			

Likelihood

Denote dataset generated by distribution with parameter θ

Likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Y}) = P(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} P(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)}|\boldsymbol{\theta})$$

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Y}) \longrightarrow \max_{\boldsymbol{\theta}}$$

samples should be i.i.d.

equivalent to

$$\log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Y}) = \sum_{i=1}^{n} \log P(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)}|\boldsymbol{\theta}) \longrightarrow \max_{\boldsymbol{\theta}}$$