CS5014 Machine Learning

Lecture 13 Unsupervised Learning

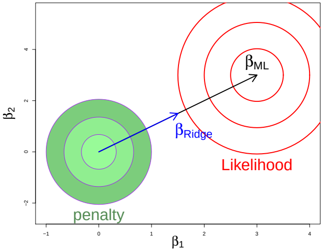
Lei Fang

School of Computer Science, University of St Andrews

Spring 2021

University of St Andrews

Some responses: geometry of ridge regression



$$\beta_{\text{ridge}} \equiv \underset{\beta}{\operatorname{argmin}} \underbrace{||\boldsymbol{y} - \boldsymbol{X}\beta||_2^2}_{L(\theta)} + \lambda||\beta||_2^2$$

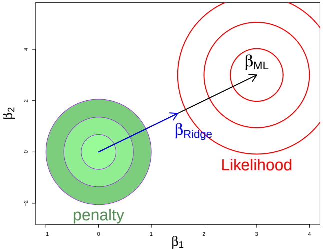
$$\beta_{\text{ridge}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{y}$$

- assume feature vectors of $\boldsymbol{X}$ are norm 1 and orthogonal i.e. orthogonormal: $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$

$$\beta_{\text{ridge}} = \frac{1}{1+\lambda}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y} = \frac{1}{1+\lambda}\beta_{ML}$$

- why the likelihood (red) contours are circular for this case ?

Some responses: geometry of ridge regression



$$\beta_{\text{ridge}} \equiv \underset{\beta}{\text{argmin}} \underbrace{||\boldsymbol{y} - \boldsymbol{X}\beta||_2^2}_{L(\theta)} + \lambda||\beta||_2^2$$

$$\beta_{\text{ridge}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{y}$$

- assume feature vectors of $\boldsymbol{X}$ are norm 1 and orthogonal i.e. orthogonormal: $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$

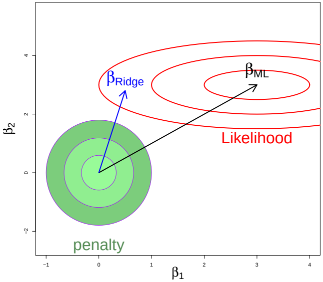
$$\beta_{\text{ridge}} = \frac{1}{1+\lambda}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y} = \frac{1}{1+\lambda}\beta_{ML}$$

- why the likelihood (red) contours are circular for this case ? remember the Hessian of $L(\theta)$?

$$H_L = 2\boldsymbol{X}^T\boldsymbol{X} \propto \boldsymbol{I}$$

- uniform discount: β_k receives the same discount: $\frac{1}{1+\lambda}$

Some responses: some other cases



When $H_L \neq \boldsymbol{I}$, the shrinkage is **NOT** uniform;

$$\beta_k^{\text{ridge}} = \frac{\sigma_k}{\sigma_k + \lambda} \beta_k^{ML}$$

- σ_k is the directional curvature of $\boldsymbol{X}^T\boldsymbol{X}$ (also the eigen value)
- flat curve (β_1 direction) $\Rightarrow$ σ_k smaller $\Rightarrow$ more discount
- curvy curve (β_2 direction) $\Rightarrow$ σ_k larger $\Rightarrow$ less discount
- makes perfect sense!
  - flat means less confident (or large variance): shrink more
  - peak means confident estimate (small variance): shrink less

($*$) the equation is true when $\boldsymbol{X}^T\boldsymbol{X}$ is a diagonal matrix; up to basis translation for more general cases

Today's topic

Unsupervised learning

- clustering
- k-means

Revisit multivariate Gaussian

Revisit k-means

- mixture of Gaussians
- EM algorithm for mixture model
  - K-means is just a specific case
- other kinds of mixture models

Unsupervised learning

Clustering

K-means

K-means

Demonstration

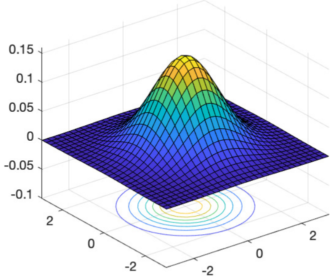
Limitations of K-means

Dissect multivariate Gaussians

$$p(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}\underbrace{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}_{d_{\boldsymbol{\Sigma}}}\right]$$

- a distance measure : (aka mahalanobis distance)

$$d_{\boldsymbol{\Sigma}}(\boldsymbol{x}; \boldsymbol{\mu}) : \text{between } \boldsymbol{x} \text{ and } \boldsymbol{\mu}$$



Dissect multivariate Gaussians

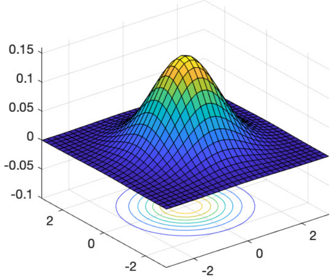
$$p(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[- \frac{1}{2} \underbrace{(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}_{d_{\boldsymbol{\Sigma}}} \right]$$

- a distance measure :(aka mahalanobis distance)

  $$d_{\boldsymbol{\Sigma}}(\boldsymbol{x}; \boldsymbol{\mu}) : \text{between } \boldsymbol{x} \text{ and } \boldsymbol{\mu}$$

- $-$: p is **negatively** related to the distance

  $$\text{larger } d_{\boldsymbol{\Sigma}}(\boldsymbol{x}; \boldsymbol{\mu}) \Rightarrow \text{further away } \boldsymbol{x} \text{ from } \boldsymbol{\mu} \Rightarrow \text{smaller } p$$



Dissect multivariate Gaussians

$$p(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}\underbrace{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}_{d_{\boldsymbol{\Sigma}}}\right]$$

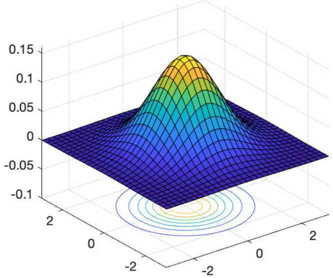
- a distance measure : (aka mahalanobis distance)

  $$d_{\boldsymbol{\Sigma}}(\boldsymbol{x}; \boldsymbol{\mu}) : \text{between } \boldsymbol{x} \text{ and } \boldsymbol{\mu}$$

- $-$: p is **negatively** related to the distance

  larger $d_{\boldsymbol{\Sigma}}(\boldsymbol{x}; \boldsymbol{\mu}) \Rightarrow$ further away $\boldsymbol{x}$ from $\boldsymbol{\mu} \Rightarrow$ smaller p

- exp: makes sure $p(\boldsymbol{x}) > 0$



Dissect multivariate Gaussians

$$p(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \boxed{\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}} \exp\left[\boxed{-}\frac{1}{2}\underbrace{\boxed{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}}_{d_{\boldsymbol{\Sigma}}}\right]$$

- a distance measure : (aka mahalanobis distance)

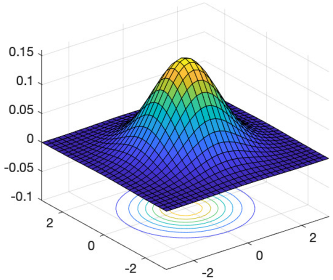
  $$d_{\boldsymbol{\Sigma}}(\boldsymbol{x}; \boldsymbol{\mu}) : \text{between } \boldsymbol{x} \text{ and } \boldsymbol{\mu}$$

- $-$: p is **negatively** related to the distance

  larger $d_{\boldsymbol{\Sigma}}(\boldsymbol{x}; \boldsymbol{\mu}) \Rightarrow$ further away $\boldsymbol{x}$ from $\boldsymbol{\mu} \Rightarrow$ smaller p

- exp: makes sure $p(\boldsymbol{x}) > 0$
- normalising constant: s.t. $\int N(\boldsymbol{x}; \cdot, \cdot) d\boldsymbol{x} = 1$;

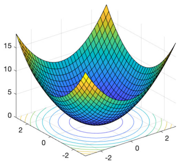
  $|\boldsymbol{\Sigma}|$: determinant; a volume measure-ish quantity

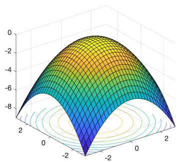


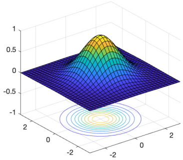
Dissect multivariate Gaussians

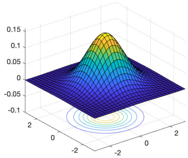
Key message: d_Σ (the distance) determines equal $p(\boldsymbol{x})$ levels

$$p(\boldsymbol{x}) \equiv \underbrace{\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}}_{C:\text{normalising cst.}} \overbrace{\exp\Big[\overbrace{-\frac{1}{2}\underbrace{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}_{f_1=d_{\boldsymbol{\Sigma}}(\boldsymbol{x};\boldsymbol{\mu})}}^{f_2}\Big]}^{f_3}$$









1. a distance measure: $f_1(\boldsymbol{x}) = d_{\boldsymbol{\Sigma}}(\boldsymbol{x};\boldsymbol{\mu})$
2. negated distance: $f_2(\boldsymbol{x}) = -\frac{1}{2}f_1(\boldsymbol{x})$
3. exp. to make sure $p > 0$: $f_3(\boldsymbol{x}) = e^{f_2(\boldsymbol{x})}$
4. scaled to make sure $\int p(\boldsymbol{x})d\boldsymbol{x} = 1$: $p(\boldsymbol{x}) = C \cdot f_3(\boldsymbol{x})$

Covariance matrix and distance

$$\boldsymbol{\Sigma}: \text{ variance-covariance matrix}$$

- $d \times d$ symmetric matrix:

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T$$

- positive definite (P.D.):

$$\boldsymbol{v}^T \boldsymbol{\Sigma} \boldsymbol{v} > 0, \quad \forall \boldsymbol{v} \in R^d$$

- why P.D. ? distance has to be positive ! (similar to univariate Gaussian: $(x - \mu)^2 \cdot \sigma^{-2} > 0$)

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) > 0, \quad \text{where } \boldsymbol{v} = \boldsymbol{x} - \boldsymbol{\mu}$$

  - if $\boldsymbol{\Sigma}$ is P.D., then $\boldsymbol{\Sigma}^{-1}$ is also P.D.; so the above is a valid distance metric

Proof: Let $\boldsymbol{y} = \boldsymbol{\Sigma} \boldsymbol{v}$; then $\boldsymbol{y}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y} = \boldsymbol{v}^T \boldsymbol{\Sigma}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{v} = \boldsymbol{v}^T \boldsymbol{\Sigma}^T \boldsymbol{v} = \boldsymbol{v}^T \boldsymbol{\Sigma} \boldsymbol{v} > 0$

Diagonal $\mathbf{\Sigma}$: implies independence

If

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_d^2 \end{bmatrix}; \quad \mathbf{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_d^2} \end{bmatrix}$$

Then

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\} = \frac{1}{(2\pi)^{d/2}(\prod_{i=1}^d \sigma_i^2)^{1/2}} \exp\{-\frac{1}{2}\sum_{i=1}^d (x_i-\mu_i)^2/\sigma_i^2\}$$

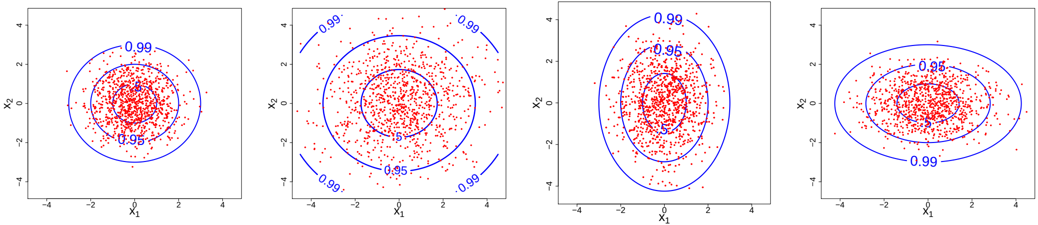
$$= \prod_{i=1}^d \underbrace{\frac{1}{(2\pi)^{1/2}\sigma_i} \exp\{-\frac{1}{2}(x_i-\mu_i)^2/\sigma_i^2\}}_{\text{unvariate Gaussian}} = \underbrace{\prod_{i=1}^d p(x_i)}_{\text{independence !}}$$

and each $p(x_i) = N(x_i; \mu_i, \sigma_i^2)$ is a univariate Gaussian

Remember independence ? it means knowing one does not inform the other: $p(x_i|\boldsymbol{x}_{/i}) = p(x_i)$

Diagonal $\boldsymbol{\Sigma}$: axis aligned ellipses

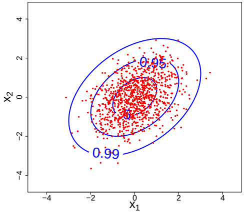
$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix};\ \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix}$$ so $d_{\boldsymbol{\Sigma}}(\boldsymbol{x}; \mathbf{0})$ are axis aligned ellipses

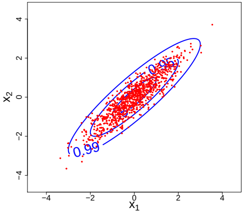


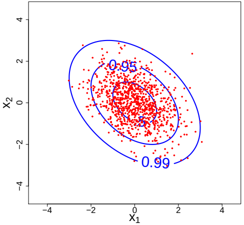
$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

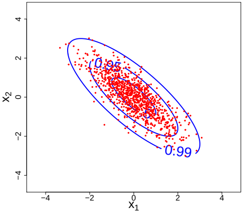
General $\boldsymbol{\Sigma}$: rotated ellipses

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \qquad d_{\boldsymbol{\Sigma}}(\boldsymbol{x}; \boldsymbol{0}) \text{ are rotated ellipses}$$









$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.4 \\ -0.4 & 1 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$$

MLE of multivariate Gaussian

Given $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}\}$, assume $\boldsymbol{x}^{(i)} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; the goal is to estimate

$$\boldsymbol{\mu},\ \boldsymbol{\Sigma}$$

The log likelihood is:

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log P(\{\boldsymbol{x}^{(i)}\}_1^m | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{m} \log N(\boldsymbol{x}^{(i)}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

The MLE is defined as usual:

$$\boldsymbol{\mu}_{ML}, \boldsymbol{\Sigma}_{ML} = \operatorname*{argmax}_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

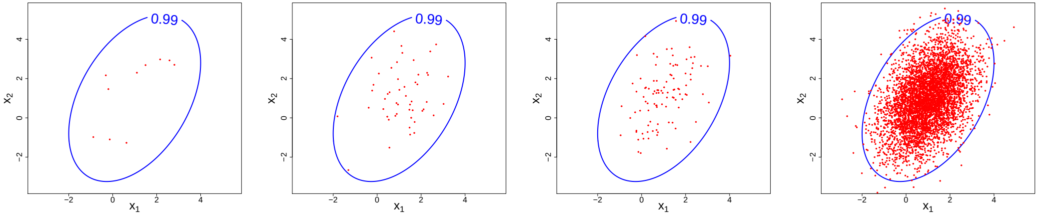
Take derivative and set to zero; after some tedious steps, the solution is:

$$\boldsymbol{\mu}_{ML} = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{x}^{(i)}, \quad \boldsymbol{\Sigma}_{ML} = \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{ML})(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{ML})^T$$

Example: MLE of MV Gaussian

True parameters: $\boldsymbol{\mu} = [1, 1]^T, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 2 \end{bmatrix}$

- MLE is **consistent**: it converges to the truth with enough data



(sample size) : 10

$\boldsymbol{\mu}_{ML} = [0.92, 1.39]^T$

$\boldsymbol{\Sigma}_{ML} = \begin{bmatrix} 1.54 & 1.45 \\ 1.45 & 2.86 \end{bmatrix}$

(sample size) : 50

$\boldsymbol{\mu}_{ML} = [1.16, 1.16]^T$

$\boldsymbol{\Sigma}_{ML} = \begin{bmatrix} 1.01 & 0.31 \\ 0.31 & 1.98 \end{bmatrix}$

(sample size) : 100

$\boldsymbol{\mu}_{ML} = [1.04, 1.19]^T$

$\boldsymbol{\Sigma}_{ML} = \begin{bmatrix} 0.87 & 0.55 \\ 0.55 & 2.01 \end{bmatrix}$

(sample size) : 5000

$\boldsymbol{\mu}_{ML} = [0.99, 0.99]^T$

$\boldsymbol{\Sigma}_{ML} = \begin{bmatrix} 0.98 & 0.6 \\ 0.6 & 2.05 \end{bmatrix}$

Finite mixture model

EM for mixture of Gaussians

Revisit K-means

Demonstration

How to decide K

EM for general mixture

EM as a general algorithm

Review: expectation

Expection of a r.v. is defined as

$$\mathbb{E}[g(X)] = \sum_x g(x)P(x) \text{ or } \mathbb{E}[g(X)] = \int g(x)P(x)dx$$

- $\mathbb{E}[a] = a$ (a is a constant)
- linearity: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$
- $\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X]$: as $\mathbb{E}[X]$ is a constant (the randomness has been integrated out)

Interpretation of Expectation: sample mean of a very large sample

$$\mathbb{E}[g(X)] = \frac{1}{m}\sum_{i=1}^{m} g(x^{(i)}); \quad m \rightarrow \infty$$

- limit of the sample average of $\{x^{(1)}, \ldots, x^{(m)}\}$ and $x^{(i)} \sim P(X)$

Review: varaiance covariance

Variance of a r.v. is defined as

$$\mathrm{Var}[g(X)] = \mathbb{E}[(g(X) - \mathbb{E}[g(X)])^2] = \mathbb{E}[g(X)^2] - \mathbb{E}[g(X)]^2$$

- $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$
- measures the spread of the distribution around the mean $\mathbb{E}[g(X)]$

Example

X is a Bernoulli r.v. with parameter $p = 0.5$; what is $\mathbb{E}[X]$?

- $\mathbb{E}[X] = 1 \times P(X = 1) + 0 \times P(X = 0) = p = 0.5$;

Y is a Binomial r.v. with $N = 10, p = 0.5$, what is $\mathbb{E}[Y]$?

- $Y = \sum_{i=1}^{N} X = N \times X$
- $\mathbb{E}[Y] = \mathbb{E}[N \times X] = N \times \mathbb{E}[X] = N \times p = 5$
- interpretation: you expect to see 5 successes out of 10 (on average the result is 5 if you repeat the experiment a lot of times)