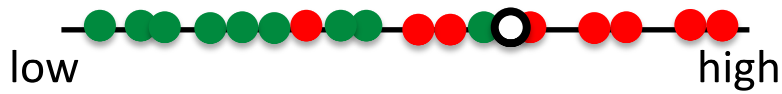
Learning Distributions
Maximum Likelihood Estimate (MLE)
Bayes Classifier

Aarti Singh

Machine Learning 10-315
Sept 14, 2020

Modeling class distribution



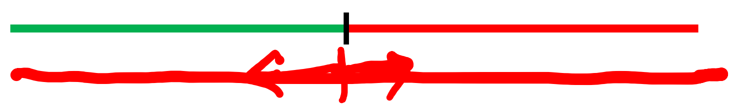
● Stress

● No Stress

○ Test subject

X, average brain activity in "Amygdala"

f(X)



$\arg\max_{y} P(X|Y)\,P(Y)$

$= \arg\max_{y} P(Y|X)$

Modeling Class distribution P(Y=y)

= Bernoulli(θ)

$P(Y = \text{●}) = \underline{\theta}$ $\qquad P(Y = \text{●}) = 1 - \theta$

Like a coin flip

Bernoulli distribution

Data, D = 

x_1 x_2 x_3 . . . .

- P(Heads) = θ, P(Tails) = 1-θ
- Flips are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Bernoulli distribution

$$P(D) = P(X_1, \ldots X_n) = \prod_{i=1}^{n} P(X_i) \equiv P_\theta(D) = P(D|\theta)$$

$$X_i \sim Ber(\theta)$$

Choose θ that maximizes the probability of observed data aka Likelihood

Maximum Likelihood Estimation (MLE)

Choose θ that maximizes the probability of observed data (aka likelihood)

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

MLE of probability of head:

heads in D

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

= 3/5

tails in D

"Frequency of heads"

Derivation

$D = \{x_1 \ldots x_n\}$

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$\arg\max_{\theta} f(\theta)$

$= \arg\max_{\theta} \log f(\theta)$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(x_i \mid \theta)$$

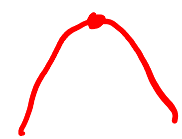
x_i ind

$$= \arg\max_{\theta} \prod_{i=1}^{\alpha_H} \theta \prod_{j=1}^{\alpha_T} (1-\theta)$$

$x_i \sim \mathrm{Ber}(\theta)$

$P(x_i = H) = \theta$

$$= \arg\max_{\theta} \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$



$$J(\theta) = \log\left(\theta^{\alpha_H}(1-\theta)^{\alpha_T}\right) = \alpha_H \log\theta + \alpha_T \log(1-\theta)$$

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{\alpha_H}{\theta} + \frac{\alpha_T}{1-\theta}(-1) = \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} \Big|_{\hat{\theta}_{MLE}} = 0$$

Derivation

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$\frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} = 0 \Rightarrow \frac{\alpha_H}{\theta} = \frac{\alpha_T}{1-\theta}$$

$$\Rightarrow \alpha_H - \alpha_H\theta = \alpha_T\theta \Rightarrow \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H+\alpha_T}$$

$X \sim Ber(\theta)$

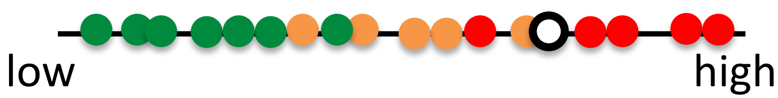
$x_1 \ldots x_n \overset{iid}{\sim} Ber(\theta)$

$P(x_i) = P(X = x_i)$

$X_i^{(j)}$ ← j^{th} feat

X_i ← data point

Modeling class distribution



X, average brain activity in "Amygdala"

f(X)

- High Stress
- Moderate Stress
- Low Stress
- Test subject

➢ How do we model multiple (>2) classes?

Modeling Class distribution P(Y) = Multinomial(p_H,p_M,p_L)

P(Y=

$P(Y = \bullet) = p_H$ $\quad P(Y = \bullet) = p_M$ $\quad P(Y = \bullet) = p_L$

Like a dice roll

$p_H + p_M + p_L = 1$

Multinomial distribution

Data, D = rolls of a dice

1, 6, 5, 2, 2, 1, 3, 4,

- $P(1) = p_1$, $P(2) = p_2$, …, $P(6) = p_6$ $\quad p_1+\ldots+p_6 = 1$
- Rolls are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Multinomial(θ) distribution where

$$\theta = \{p_1, p_2, \ldots, p_6\}$$

<u>Choose θ that maximizes the probability of observed data aka "Likelihood"</u>

Maximum Likelihood Estimation (MLE)

Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$= p_1 \dots p_6$

$p_1 + \dots + p_6 = 1$

MLE of probability of rolls:

$$\widehat{\theta}_{MLE} = \hat{p}_{1,MLE}, \dots, \hat{p}_{6,MLE}$$

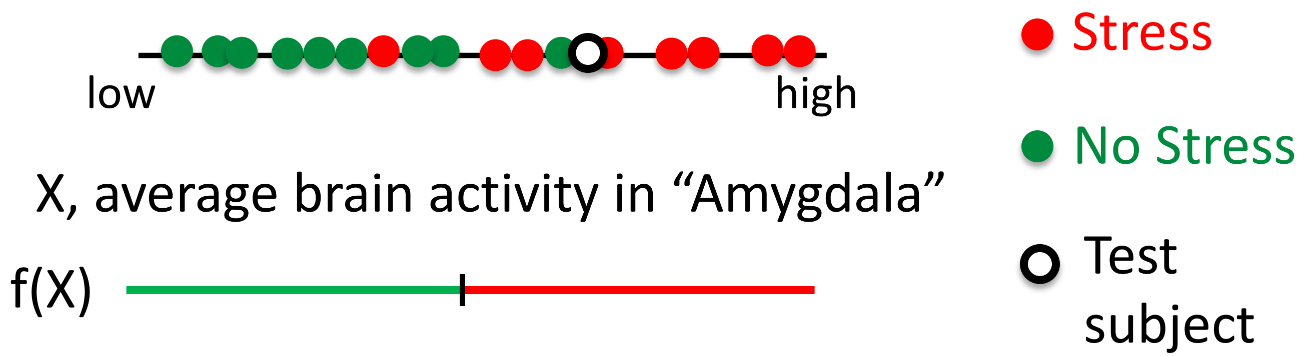
$$\hat{p}_{y,MLE} = \frac{\alpha_y}{\sum_y \alpha_y}$$

α_y ← # times dice rolls y ← Rolls that turn up y

$\sum_y \alpha_y$ ← total # rolls ← Total number of rolls

"Frequency of roll y"

Bayes Classifier



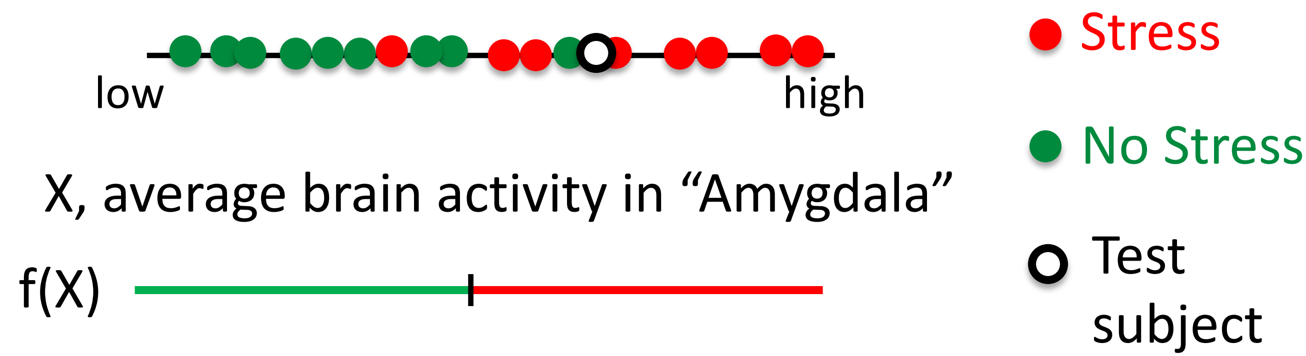
$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\substack{\text{Class conditional} \\ \text{Distribution of features}}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

We can now consider appropriate distribution models for the two terms:

Class distribution P(Y=y)

Class conditional distribution of features P(X=x|Y=y)

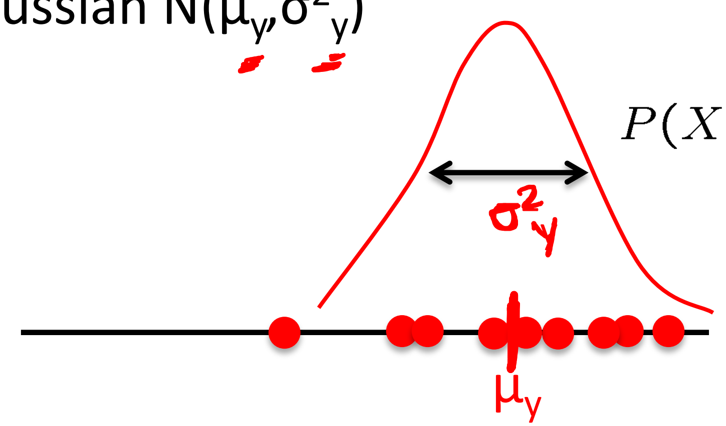
Modeling class conditional distribution of features



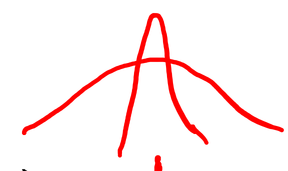
Modeling class conditional distribution of feature P(X=x|Y=y)

- What distribution would you use?

E.g. P(X=x|Y=y) = Gaussian N(μ_y,σ^2_y)



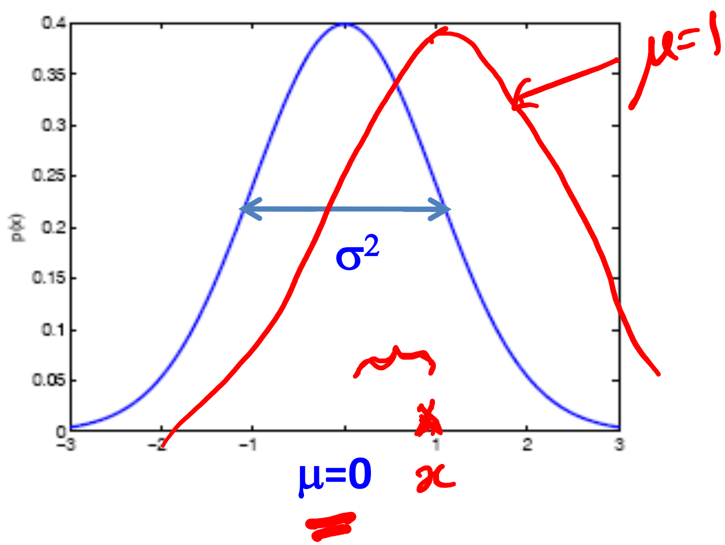
$$P(X = x|Y = \bullet)$$

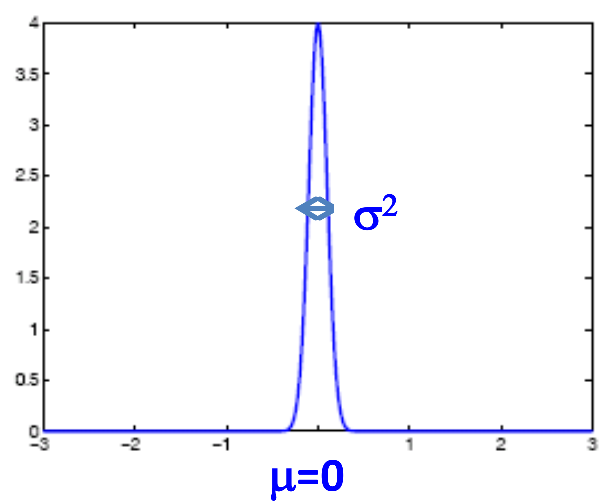


1-dim Gaussian distribution

X is Gaussian $N(\mu,\sigma^2)$

$$P(X = x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$





Why Gaussian?

- Properties μ, Σ $E[X_i^2]$ $E[X_i X_j]$
  - Fully Specified by first and second order statistics
    - Uncorrelated ⇔ Independence $P(X,Y) = P(X)P(Y)$
  - X, Y Gaussian => aX+bY Gaussian $E[X,Y] = 0$

  - <u>Central limit theorem:</u> if $X_1, \ldots, X_n$ are any iid random variables with mean μ and variance $\sigma^2 < \infty$

    then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow[n\to\infty]{} N(0,\sigma^2)$$

d-dim Gaussian distribution

$X = \begin{bmatrix} X_1 \\ \vdots \\ X_d \end{bmatrix} \leftarrow$ brain activity @ pixel i

X is Gaussian N(μ, Σ) μ is d-dim vector, Σ is dxd dim matrix

$\mu_i = E[X_i]$ $\Sigma_{ji} = \Sigma_{ij} = E[(X_i - E[X_i])(X_j - E[X_j])]$

$$P(X = x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right),$$

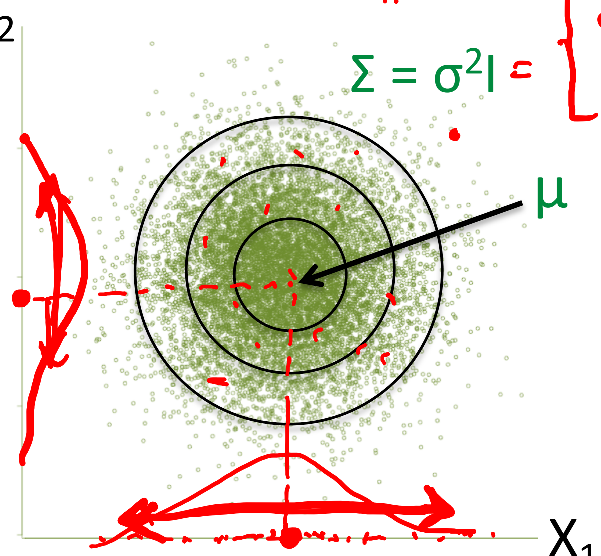
d $\frac{d(d+1)}{2}$

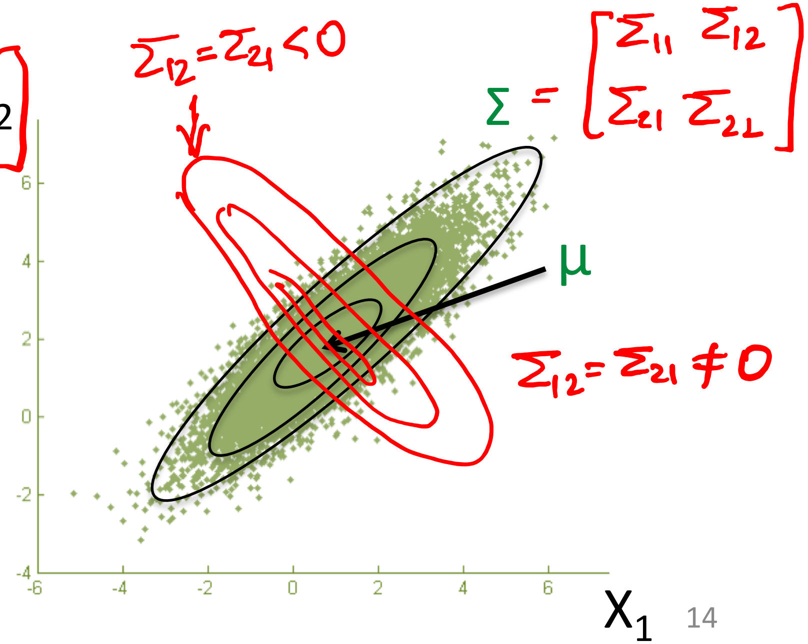
Σ_{11}?

d=2

$X = [X_1; X_2]$

$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ $\hookrightarrow E[X_2^2]$





How to learn parameters from data? MLE

(Continuous case)

Gaussian distribution

Data, D =

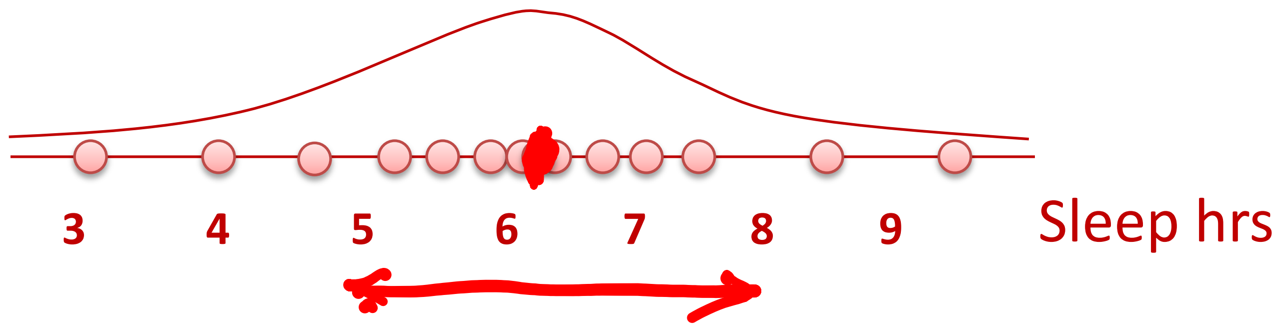
3 4 5 6 7 8 9 Sleep hrs

How many hours did you sleep last night?

- Poll

Gaussian distribution

Data, D =



- Parameters: μ – mean, σ^2 - variance

x_i – sleep hrs of person i
– avg brain activity of person i

- Sleep hrs are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Gaussian distribution

Maximum Likelihood Estimation (MLE)

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$D = \{X_1 \ldots X_n\}$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

Maximum Likelihood Estimation (MLE)

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i | \theta)$$

Independent draws

$= \mu, \sigma^2$

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \underline{e^{-(X_i-\mu)^2/2\sigma^2}}$$

Identically distributed

Maximum Likelihood Estimation (MLE)

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$ Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X_i-\mu)^2/2\sigma^2}$$ Identically distributed

$$= \arg \max_{\theta=(\mu,\sigma^2)} \underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}}_{J(\theta)}$$

$$= \arg\min_{\mu} \overbrace{\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\cancel{2\sigma^2}}}^{J(\mu)}$$

$$\frac{\partial}{\partial \mu} J(\mu) = 0$$

Derivation

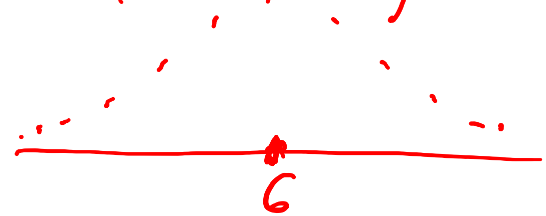
$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

- Breakout

Groups 1-10: Jamboard_1_10
Groups 11-20: Jamboard_11_20

MLE for Gaussian mean and variance

$X_1 .. X_n \sim N(\mu, \sigma^2)$



$E[\hat{\mu}_{MLE}] = \mu$ ✓

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \underbrace{\hat{\mu}}_{=\hat{\mu}_{MLE}})^2$$

$E[\hat{\sigma}^2_{MLE}] \neq \sigma^2$

$\frac{n-1}{n}\sigma^2$

(unbiased) $n-1$

Gaussian Bayes classifier

cont.

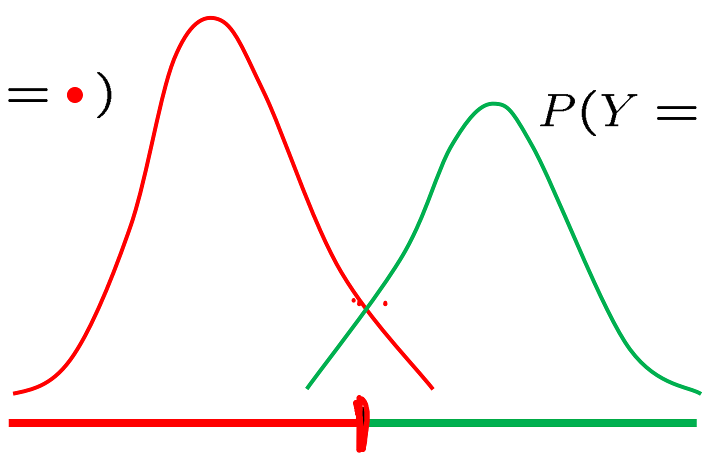
$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

How to learn parameters θ, μ_y, Σ_y from data?

Class conditional
Distribution of features → Gaussian(μ_y, Σ_y)

Class distribution → Bernoulli(θ)

$P(Y = \bullet)P(X = x|Y = \bullet)$



$P(Y = \bullet)P(X = x|Y = \bullet)$

1-dim Gaussian Bayes classifier

$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

Class conditional Distribution of features → Gaussian(μ_y, σ^2_y)

Class distribution → Bernoulli(θ)

- What decision boundaries can we get in 1-dim?

$P(Y = \bullet)P(X = x|Y = \bullet)$

$P(Y = \bullet)P(X = x|Y = \bullet)$