Lecture Notes 01/25/2021

Admin

- HW #1 due $\boxed{01/29/2021}$ @ 11:30 PM — Upload to D2L
  - → Submit theory as a pdf (convert via word, camscan, etc)
  - → upload code as a pdf
    - Show outputs / figures
    - Convert Jupyter notebooks to pdf

0 1 2

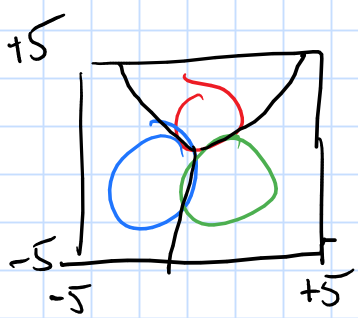
$p = [0.5 \ .2 \ .3]$

Homework help

$g_i(x)$ → param: μ_i, Σ_i, $P(\omega_i)$ $\propto$ Posterior

D_{Tr} → Learn μ_1, μ_2, μ_3, $\Sigma_1 \ \Sigma_2 \ \Sigma_3$

D_{TE}



1×1

$[1\times d] \quad [d\times d] \quad [d\times 1]$

$$g_i(x) = -\frac{1}{2}\underbrace{(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)} - \boxed{\frac{d}{2}\log(2\pi)} - \frac{1}{2}\log(|\Sigma_i|) + \log p(\omega_i)$$

Assumed column vectors

2D
100 samples

$D = (100\times 2)$

D[0].T np.linalg.transpose(x)

$N \in \{0, 1, 2\}$

$N \to 2$

$p_0, p_1, p_2 \to p_1 + p_2 + p_3 = 1$ ✩

$p_1, p_2, p_3 \in [0, 1]$

$p_0 \downarrow 0.7$, $p_1 \downarrow 0.25$, $p_2 \downarrow 0.05$

M

0.7 0.95 1

$r = \text{sample}(1000, [p_0, p_1, p_2])$

0 1 2

- Look at the CDF
- $\eta \sim \text{Uni}(0, 1)$

$0.6 \to 0$

$0.9 \to 1$

$r \in N^{1000}$

p

$c \to \text{cdf}$

```
for m in range(M):
    etu ← Sample Uni(0,1)
    L ← c ≤ etu
    U ← c > etu
```

$g(x) = \bar{w}^T x + w_0$

Goal: Find $\bar{w}$ and w_0.

$\bar{w} \downarrow$ weight vector

$w_0 \downarrow$ bias

$$w = \begin{bmatrix} w_0 \\ \bar{w} \end{bmatrix}$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

$$\ell(w) = \sum_{i=1}^{n} (\underbrace{y_i}_{\text{target}} - \underbrace{w^T x_i}_{\text{prediction}})^2 = (Y - Xw)^T (Y - Xw)$$

$$\underset{n \times d}{X} \; \underset{d \times 1}{w}$$

$$n \times 1 - \underbrace{n \times 1}$$

$\boxed{\text{We want the } w \text{ that minimizes the loss function wrt } w!}$

$$\frac{\partial L}{\partial w} = 2X^T(y - Xw) = 0$$

$\Rightarrow$ Solve for w

$$w = (X^T X)^{-1} X^T y$$

Something to think about

- $X^T X$ must be invertible
- What if n and d are really large

How do we implement this?

(a) targets $y_i \in \{\pm 1\}$ $y_i = 1.29$ [training]

(b) $\hat{w} = (X^T X)^{-1} X^T y$

(c) $\hat{y} = w^T x \rightarrow x$ is not from the training data