3.

Review

Logistic Regression

Given $D = \{(x_i, y_i)\}_{i=1}^{N}$ $(x_i \in \mathbb{R}^D,\ y_i \in \mathbb{R})$

Logistic function

$$f(x; w, b) = \frac{1}{1 + \exp(-(w^T x + b))} \in [0, 1]$$

Cross entropy loss

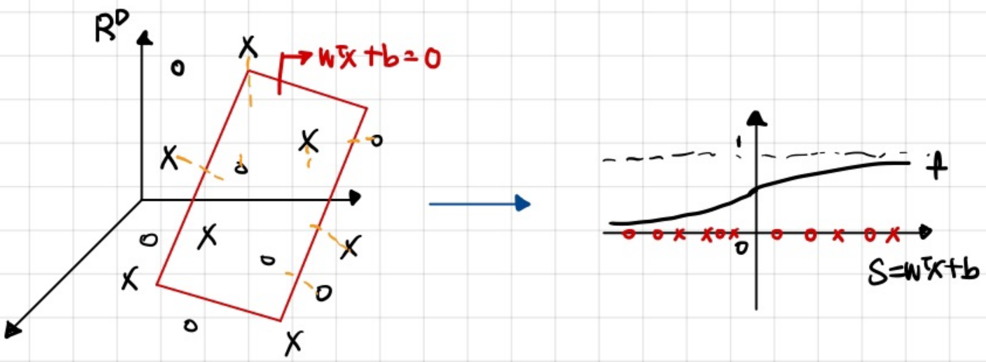
$$l(f_i, y_i) = -y_i \log f_i - (1 - y_i) \log(1 - f_i) \qquad f_i = f(x_i; w, b)$$

$$(w^*, b^*) = \underset{w,b}{\operatorname{argmin}}\ \hat{L}(w, b) \ / \ \hat{L}(w, b) = \sum_{i=1}^{N} l(f(x_i; w, b), y_i)$$

Gradient descent

$$\begin{cases} w^{(t+1)} = w^{(t)} - \varepsilon \frac{\partial L}{\partial w}\Big|_{w^{(t)}, b^{(t)}} = w^{(t)} - \varepsilon \sum_{i=1}^{N} (f_i - y_i) x_i \\ b^{(t+1)} = b^{(t)} - \varepsilon \sum_{i=1}^{N} (f_i - y_i) \end{cases}$$

when $\underset{\sim}{\varepsilon} > 0$
step size



Projection to Hyperplane

Bayesian approach

Prior : $P(w, b)$, $D = \{(x_i, y_i)\}_{i=1}^{N}$ Posterior : $P(w, b \mid D) = \frac{P(D \mid w, b) P(w, b)}{P(D)}$

Probability density function

Distribution about given data

$\propto P(w, b \mid D) P(D)$

Goal : $P(w, b)$, $P(D \mid w, b)$, $D \Rightarrow P(w, b \mid D)$

Likelihood distribution

For logistic regression

→ 각 data끼리는 independent 하다는 가정

$w, b \,/\, x \,/\, y$

$$P(D|w,b) = \prod_{i=1}^{N} P(y_i|x_i,w,b)\,P(x_i)$$ (x_i와 $\theta(w,b)$는 독립?)

$$P(y_i|x_i,w,b) = \frac{1}{1+\exp(-(w^T x+b))}$$ [$P(w,b)$ is chosen by human]

Posterior / Prior / likelihood = $P(D|w,b)$

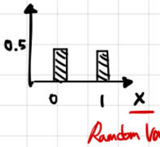
$$\Rightarrow P(w,b|x_1,y_1,\dots,x_N,y_N) \propto P(w,b) \times \prod_{i=1}^{N} P(y_i|x_i,w,b)$$

Prediction for a new x : predictive $P(y|x,D) = \int P(y|x,w,b)\cdot P(w,b|D)\,dw\,db$

Test data / Test data / Training data

Basic Probability

Ex) 동전 던지기 $H=1$, $T=0$



$P(x=0) : \frac{1}{2}$
$P(x=1) : \frac{1}{2}$

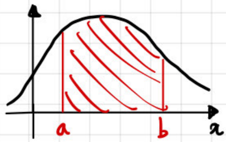
Probability mass function

· Discrete case

$P(x) \geq 0$ for $x \in \{1,2,\dots,C\}$

$\sum_{x=1}^{C} P(x) = 1$

· Continuous case

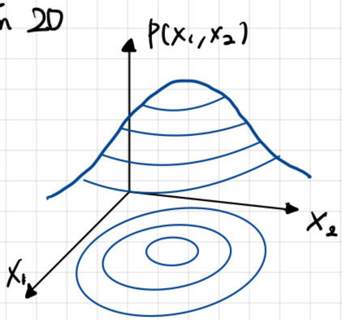


$P(x\in[a,b]) = \int_a^b P(x)\,dx$

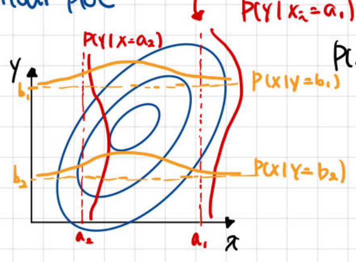
$P(x) \geq 0,\ \forall x \in \mathbb{R}$

$\int_{-\infty}^{\infty} P(x)\,dx = 1$

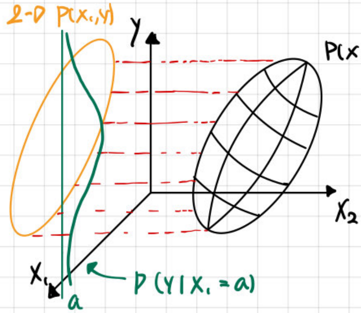
· In 2D



Contour plot



$P(x,y)$: Joint Pdf (Probability of x and y occurring jointly)



$P(x_1,x_2,y)$: Joint pdf

Marginal pdf

$- P(x_1,y) = \int P(x_1,x_2,y)\,dx_2$

Conditional Probability

$P(x,y) = P(y|x)P(x) = P(x|y)P(y)$

Bayes Rule

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

For $x \in \mathbb{R}^D$

- Expectation of x: $E_x[x] = \int x \cdot P(x)\,dx \in \mathbb{R}^D$
- Covariance matrix
  - $Cov[x] = E_x[(x - E[x])(x - E[x])^T] \in \mathbb{R}^{D\times D}$ (2-D)
    Matrix multiplication: $(D\times 1) \times (1\times D) = D\times D$
  - $(Cov[x])_{ij} = E_x[(x_i - E[x]_i)(x_j - E[x]_j)]$
- Correlation coefficient

$$\rho_{ij} = \frac{Cov[x]_{ij}}{\sqrt{Cov[x]_{ii}} \cdot \sqrt{Cov[x]_{jj}}}$$

Ex) p.d.f of Gaussian distribution for $x \in \mathbb{R}^D$

$$\boxed{P(x) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}$$

$\mu \in \mathbb{R}^D$
$\Sigma \in \mathbb{R}^{D\times D}$: symmetric, positive-definite

$\int P(x)\,dx = 1$

Positive-definite?

$x^T\Sigma x > 0$ for $\forall x \neq 0$

ex) $I_{2\times2} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $\quad x^Tx = x_1^2 + x_2^2 = 0 \iff x_1 = x_2 = 0$

$E_x[x] = \int_{\mathbb{R}^D} x P(x)\,dx = \mu$, $\quad Cov[x] = \int_{\mathbb{R}^D} (x-\mu)(x-\mu)^T P(x)\,dx$

$= \Sigma$ (Covariance matrix)

4.

Review

P.d.f of one-dim Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi}^{D}|\Sigma|^{\frac{1}{2}}} \exp\left(\underbrace{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}_{\text{Quadratic function respect to } x}\right) \quad \mu \in \mathbb{R}^D,\ \Sigma \in \mathbb{R}^{D\times D}$$

Σ: Symmetric and positive-definite (SPD)

$$E[x] = \int_{\mathbb{R}^D} x\,p(x)\,dx = \mu \in \mathbb{R}^D$$

$$= \int_{\mathbb{R}^D} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} p(x)\,dx = \begin{bmatrix} \int_{\mathbb{R}^D} x_1 p(x)\,dx \\ \vdots \\ \int_{\mathbb{R}^D} x_D p(x)\,dx \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$$

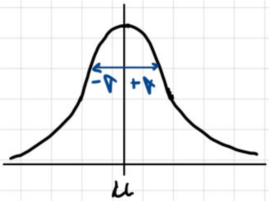
$$\mathrm{Var}[x] = E\left[(x - E(x))^2\right]$$

$$\mathrm{Cov}[x_1, x_2] = E[(x_1 - E[x_1])(x_2 - E[x_2])]$$

$$\mathrm{Cov}[x] = \int_{\mathbb{R}^D} (x-\mu)(x-\mu)^T p(x)\,dx = \Sigma \text{ (Covariance matrix) } \in \mathbb{R}^{D\times D}$$

$$\mathrm{Cov}[x]_{ij} = \int (x_i - \mu_i)(x_j - \mu_j)\,p(x)\,dx$$

$$\mathrm{Cov}[x] \in \mathbb{R}^{D\times D}\ (\Sigma) = \begin{bmatrix} \int (x_1-\mu_1)^2 p(x)\,dx & \int (x_1-\mu_1)(x_2-\mu_2)\,p(x)\,dx & \cdots & \\ \vdots & \int (x_2-\mu_2)^2 p(x)\,dx & & \\ & & \ddots & \\ & & & \int (x_D-\mu_D)^2 p(x)\,dx \end{bmatrix}$$

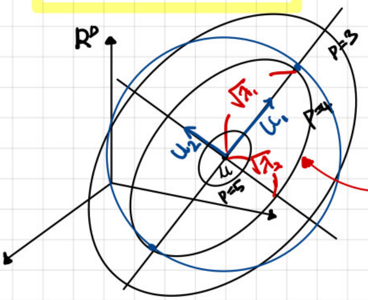


Consider the eigendecomposition of Σ

$$\boxed{\Sigma u_i = \lambda_i u_i}$$

u_i : Eigen vector, λ_i : Eigen value

$\lambda_i > 0$ (when $i = 1, \cdots, D$), $u_i \in \mathbb{R}^D$



Ellipse contour

$p(x) = p(\mu)\cdot \exp(-\frac{1}{2})$ or

$(x-\mu)^T \Sigma^{-1} (x-\mu) = 1$

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{pmatrix} \begin{matrix} = X_a \\ \\ = X_b \end{matrix} \longrightarrow X = \begin{pmatrix} X_a \\ X_b \end{pmatrix} \in \mathbb{R}^D, \ \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \in \mathbb{R}^D$$

$$\Sigma = \begin{array}{cc} \in \mathbb{R}^{D_a \times D_a} & \mathbb{R}^{D_a \times D_b} \\ \end{array}$$

$$\Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} \in \mathbb{R}^{D \times D}$$

$$\begin{array}{cc} \in \mathbb{R}^{D_b \times D_a} & \in \mathbb{R}^{D_b \times D_b} \end{array}$$

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

$\exp(x_a) \times \exp(x_b|x_a) = \exp(x_a + x_b|x_a)$

$$= \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\Big(\underbrace{-\frac{1}{2}(x_{aa}-\mu_{aa})^T \Sigma_{aa}^{-1} (x_{aa}-\mu_{aa})}_{P(x_a)} \underbrace{-\frac{1}{2}(x_b-\mu_{b|a})^T \Sigma_{b|a}^{-1} (x_b-\mu_{b|a})}_{P(x_b|x_a)}\Big)$$

$$= P(x_b|x_a)P(x_a) = P(x_a, x_b)$$

Where $\mu_{b|a} = \mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a)$, $\Sigma_{b|a} = \Sigma_{bb} - \Sigma_{ba} \cdot \Sigma_{aa}^{-1}\Sigma_{ab}$

We prove $\longrightarrow$ $P(x_a), P(x_b|x_a)$ are all Gaussians

Perfect square form

$-\frac{1}{2}x^TAx + b^Tx$ when $A = A^T$ (Symmetric)

이 식을 $(x+c)^TA(x+c)+d$ 꼴으로 만들어보자

$$= -\frac{1}{2}x^TAx + b^TA^{-1}Ax$$

$$= -\frac{1}{2}x^TAx + \frac{1}{2}b^TA^{-1}Ax + \frac{1}{2}x^TAA^{-1}b$$

$$= -\frac{1}{2}x^TA(x-A^{-1}b) + \frac{1}{2}b^TA^{-1}Ax - \frac{1}{2}b^TA^{-1}AA^{-1}b + \frac{1}{2}b^TA^{-1}AA^{-1}b$$

$$= -\frac{1}{2}x^TA(x-A^{-1}b) + \frac{1}{2}b^TA^{-1}A(x-A^{-1}b) + \frac{1}{2}b^TA^{-1}AA^{-1}b$$

$$= -\frac{1}{2}(x-A^{-1}b)^TA(x-A^{-1}b) + c$$

$$\boxed{-\frac{1}{2}x^TAx + b^Tx = -\frac{1}{2}(x-A^{-1}b)^TA(x-A^{-1}b) + c}$$

Ex) $C \cdot \exp(-(x_1^2 + 4x_2^2 + 2x_1 + 8x_2))$

$$= C \cdot \exp\left(-\frac{1}{2} x^T \underbrace{\begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix}}_{A} x - \underbrace{\begin{pmatrix} 2 \\ 8 \end{pmatrix}}_{b}^T x\right)$$

$$= C \cdot \exp\left(-\frac{1}{2} \left(x + \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{8} \end{pmatrix}\begin{pmatrix} 2 \\ 8 \end{pmatrix}\right)^T \begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix}\left(x + \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) + \underset{\sim}{c}\right)$$

$$= \underset{\sim}{C'} \cdot \exp\left(-\frac{1}{2}\left(x + \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)^T \begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix}\left(x + \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right)\right)$$

$$= C' \cdot \exp\left(-\frac{1}{2}\begin{pmatrix} x_1+1 \\ x_2+1 \end{pmatrix}^T \begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix}\begin{pmatrix} x_1+1 \\ x_2+1 \end{pmatrix}\right)$$

$$= C' \cdot \exp\left(-\frac{1}{2}(2(x_1+1) + 8(x_2+1)) \begin{pmatrix} x_1+1 \\ x_2+1 \end{pmatrix}\right)$$

$$= -(x_1+1)^2 - 4(x_2+1)^2$$

$$-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) = -\frac{1}{2}\left(\begin{pmatrix} x_a \\ x_b \end{pmatrix} - \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}\right)^T \qquad \Sigma^{-1} = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} = \Lambda$$

$$= -\frac{1}{2}\left(\begin{pmatrix} x_a \\ x_b \end{pmatrix} - \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}\right)^T \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}\left(\begin{pmatrix} x_a \\ x_b \end{pmatrix} - \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}\right)$$

$(P(x_a, x_b) = P(x_b | x_a) P(x_a))$

여기를 x_b만 포함한 식

Terms containing term x_b

→ x_a는 constant 취급

$$-\frac{1}{2} x_b^T \Lambda_{bb} x_b - x_b^T \Lambda_{ba}(x_a - \mu_a) + x_b^T \Lambda_{bb}\mu_b + C$$

$$= -\frac{1}{2}\left(x_b - (\mu_b - \Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a))\right)^T \cdot \Lambda_{bb}(x_b - (\mu_b - \Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a)))$$

$$= -\frac{1}{2}(x_b - \mu_b)^T \Lambda_{bb}(x_b - \mu_b) - (x_b - \mu_b)^T \Lambda_{ba}(x_a - \mu_a) - \frac{1}{2}(x_a - \mu_a)^T \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a)$$

Remaining terms

$$-\frac{1}{2}(x_a - \mu_a)^T \Lambda_{aa}(x_a - \mu_a) + \frac{1}{2}(x_a - \mu_a)^T \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a)$$

$$= -\frac{1}{2}(x_a - \mu_a)^T \underbrace{(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})}_{\Sigma_{a|b}}(x_a - \mu_a)$$

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix} \quad \text{when } M = (A - BD^{-1}C)^{-1}$$

Should not be remembered

(Schur's Complement)

$$\Lambda^{-1} = \Sigma \rightarrow = (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1} = \Sigma_{aa}$$