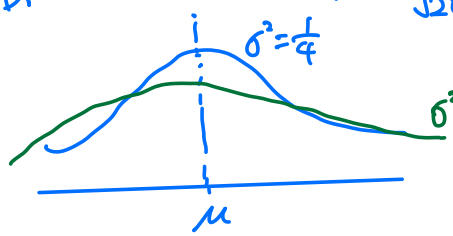
CS136 lecture 6 Gaussian model

Outline

1. Univarariate Gaussian distribution
2. MLE for gaussians
3. Biased & unbiased estimators
4. Special properties of Gaussian
5. Covariance and covariance matrices

Univarariate Gaussian distribution

$$\text{PDF}: \text{NormPDF}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$\int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$$

$$\Rightarrow \frac{1}{\sqrt{2\pi}\sigma} \int e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$$

$$\Rightarrow \int e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sqrt{2\pi}\sigma$$

Computing its mean and variance

Mean:

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} x e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \quad *$$

let $t = \frac{x-\mu}{\sigma} \Rightarrow x = \sigma t + \mu$

$dt = \frac{1}{\sigma} dx \Rightarrow dx = \sigma\, dt$

$$* = \frac{1}{\sqrt{2\pi}\cancel{\sigma}} \int_{-\infty}^{+\infty} (\sigma t + \mu) \cdot \cancel{\sigma}\, e^{-\frac{t^2}{2}} dt$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\sigma t + \mu) \cdot e^{-\frac{t^2}{2}} dt$$

$$= \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{+\infty} \sigma t \cdot e^{-\frac{t^2}{2}} dt + \int_{-\infty}^{+\infty} \mu e^{-\frac{t^2}{2}} dt \right)$$

$$= \frac{1}{\sqrt{2\pi}} \left(\sigma \underbrace{\int_{-\infty}^{+\infty} t e^{-\frac{t^2}{2}} dt}_{A} + \mu \underbrace{\int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} dt}_{B} \right)$$

$$= \frac{1}{\sqrt{2\pi}} \left(\sigma \cdot 0 + \sqrt{2\pi}\, \mu \right)$$

$$= \mu$$

A:

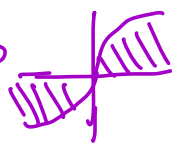
$$f(t) = \int_{-\infty}^{+\infty} t e^{-\frac{t^2}{2}} dt$$

$$f(-t) = -\int_{-\infty}^{+\infty} t e^{-\frac{t^2}{2}} dt$$

$$-f(t) = -\int_{-\infty}^{+\infty} t e^{-\frac{t^2}{2}} dt$$

odd function: $f(-x) = -f(x)$

$\int_{-\infty}^{+\infty}$ odd function $= 0$



B:

$$\int e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sqrt{2\pi}\,\sigma$$

here. $\mu = 0$. $\sigma = 1$

$$\Rightarrow \int e^{-\frac{t^2}{2}} dt = \sqrt{2\pi}$$

Variance

$$Var(X) = E(X^2) - E(X)^2$$

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}} dx \qquad **$$

let $t = \frac{x-\mu}{\sigma}$. $\Rightarrow$ $x = \sigma t + \mu$

$dt = \frac{1}{\sigma} dx$ $\qquad dx = \sigma\, dt$

$$** \Rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\sigma t + \mu)^2 \cdot e^{-\frac{t^2}{2}} dt$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\sigma^2 t^2 + 2\mu\sigma t + \mu^2) e^{-\frac{t^2}{2}} dt$$

A: $\int_{-\infty}^{+\infty} \sigma^2 t^2 e^{-\frac{t^2}{2}} dt = \sigma^2 \int_{-\infty}^{+\infty} t^2 e^{-\frac{t^2}{2}} dt = \sqrt{2\pi}\sigma^2$ $\qquad$ A

$$B: \int_{-\infty}^{+\infty} \left(2\mu\sigma t\, e^{-\frac{t^2}{2}} dt \right) = 2\mu\sigma \left(\int_{-\infty}^{+\infty} t\, e^{-\frac{t^2}{2}} \right) = 0$$

$$C: \int_{-\infty}^{+\infty} \mu^2 e^{-\frac{t^2}{2}} dt = \mu^2 \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} dt = \sqrt{2\pi}\,\mu^2$$

$$\Rightarrow \frac{1}{\sqrt{2\pi}} \left(\sqrt{2\pi}\,\sigma^2 + \sqrt{2\pi}\,\mu^2 \right) = \sigma^2 + \mu^2$$

$$\Rightarrow E(X^2) = \sigma^2 + \mu^2$$

Besides. $E(X)^2 = \mu^2$

$\therefore$ $Var(X) = \sigma^2$

MLE for gaussians / biased & unbiased estimators

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

MLE: $f(y; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

for μ:

$$f(y;\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$= \sum_{i=1}^{n} -(\log\sqrt{2\pi} + \log\sigma) - \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial f(y;\mu)}{\partial \mu} = -2 \sum_{i=1}^{n} (x_i - \mu) \overset{\text{set}}{=} 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i - n\mu = 0$$

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

$$E(\hat{\mu}) = E\left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)$$

$$= \frac{1}{n} E\left(\sum_{i=1}^{n} x_i \right)$$

$$= \frac{1}{n} \cdot n \cdot \mu$$

$$= \mu \quad \rightarrow \text{unbiased estimator}$$

$$f(y;\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$= \sum_{i=1}^{n} -(\cancel{\log\sqrt{2\pi}} + \log\sigma) - \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$= -\sum_{i=1}^{n} \log\sigma + \frac{(x_i-\mu)^2}{2} \cdot \frac{1}{\sigma^2}$$

$$\frac{\partial f(y;\sigma)}{\partial \sigma} = -\sum_{i=1}^{n} \left(\frac{1}{\sigma} - (x_i-\mu)^2 \cdot \frac{1}{\sigma^3} \right) \overset{\text{set}}{=} 0$$

$$\times \sigma^3 \Rightarrow \sum_{i=1}^{n} \left(\sigma^2 - (x_i-\mu)^2 \right) = 0$$

$$n\sigma^2 - \sum_{i=1}^{n} (x_i-\mu)^2 = 0$$

$$\boxed{\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i-\mu)^2}$$ ← biased estimation

$$E(\sigma^2) = E\left[\frac{1}{n} \sum_{i=1}^{n} (x_i-\hat{\mu})^2\right]$$

$$= \frac{1}{n} E\left[\sum_{i=1}^{n} (x_i-\hat{\mu})^2\right]$$

$$= \frac{1}{n} E\left[\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\hat{\mu} + \sum_{i=1}^{n} \hat{\mu}^2\right]$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\frac{1}{n}\sum_{i=1}^{n} x_i\,\hat{\mu} + \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}\cdot\hat{\mu}\right)$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\hat{\mu}^2 + \hat{\mu}^2\right)$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \hat{\mu}^2\right)$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2 - \hat{\mu}^2 + \mu^2\right]$$

$$= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2\right) - \left(\mu^2 - \hat{\mu}^2\right)\right]$$

$$= \left[\underbrace{E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2\right)}_{A} - \underbrace{E\left(\hat{\mu}^2 - \mu^2\right)}_{B}\right] \quad ****$$

A: $E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2\right)$

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\mu)^2$$

$$= \frac{1}{n} \cdot n \cdot \sigma^2$$

$$= \sigma^2$$

B: $E(\hat{\mu}^2 - \mu^2)$

$= E(\hat{\mu}^2) - \mu^2$ ← $E(\hat{\mu}) = \mu$

$= E(\hat{\mu}^2) - E(\hat{\mu})^2$ but not $\hat{\mu} \neq \mu$

$= \mathrm{Var}(\hat{\mu})$ not $E(\mu^2) = \mu^2$

$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$

$= \frac{1}{n^2} \cdot n \cdot \sigma^2$

$= \frac{\sigma^2}{n}$

$\Rightarrow \sigma^2 + \frac{\sigma^2}{n} = \frac{n+1}{n}\sigma^2 \neq \sigma^2 \Rightarrow$ biased estimator.

To correct:

$\frac{N}{N+1}\sigma^2$ → unbiased estimator!

Useful properties of Gaussian

Assume $X \sim N(\mu_X, \sigma_X^2)$

$Y \sim N(\mu_Y, \sigma_Y^2)$

① Gaussian 的线性变换还是 Gaussian

② 两个 Gaussian 相加还是 Gaussian

③ 两个 Gaussian 相乘是 Gaussian

then.

① Let $Z = aX + b$, then $Z \sim N(a\mu_X + b, a^2\sigma_X^2)$ (Linear transformer)

② Let $S = X + Y$, then $S \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$ (Sum and product)

③ Products of Gaussian r.v are NOT Gaussian.

However, the product of Gaussian PDFs has Gaussian form.

$\mathrm{NormPDF}(x \mid \mu_X, \sigma_X^2) \cdot \mathrm{NormPDF}(x \mid \mu_Y, \sigma_Y^2) \propto \mathrm{NormPDF}(x \mid \mu, \sigma^2)$

where $\mu = \frac{\sigma_X^2\mu_X + \sigma_Y^2\mu_Y}{\sigma_X^2 + \sigma_Y^2}$ $\sigma^2 = \frac{1}{\frac{1}{\sigma_X^2} + \frac{1}{\sigma_Y^2}}$

Covariance and covariance Matrices

$$\mathrm{Cov}(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

If X is d dimensional random vector, its covariance matrix is a **symmetric positive definite matrix**

$$\mathrm{Cov}(X) = E[(X - E(X))(X - E(X))^T]$$

$$= \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & & \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_{d-1}, X_d) \\ & & \mathrm{Cov}(X_d, X_{d-1}) & \mathrm{Var}(X_d) \end{pmatrix}$$

Review: Positive Definite matrices (Gilbert P415)

Some properties:

① $\lambda_i > 0$

② $X^T S X$ is positive for all vector $x \neq 0$

Ex:

Energy $x^T S x = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 4 & 9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2x_1^2 + 8x_1x_2 + 9x_2^2$

Is this positive for every x_1 and x_2 except $(x_1, x_2) = (0, 0)$? *Yes, it is a sum of squares*:

$$x^T S x = 2x_1^2 + 8x_1x_2 + 9x_2^2 = 2(x_1 + 2x_2)^2 + x_2^2 = \text{positive energy.}$$

③ $S = A^T A$ for matrix A with independent columns

④ All the leading determinants $D_1, D_2, \cdots, D_n$ of S are positive

CS136 Lecture 7

Bishop 2.3.1 – 2.3.6

Outline

1. Multivariate Gaussian Distribution
2. Covariance Properties: Why Contours are Elliptical
3. Margins of Gaussian are Gaussian
4. Conditionals of Gaussian are Gaussian
5. Linear-Gaussian models are joint Gaussian.

Multivariate Gaussian Distribution

31λ

Joint Distribution of two independent gaussians (independent $\Rightarrow$ $Cov(X_1, X_2) = 0$)

$X_1 \sim N(\mu_1, \sigma_1^2)$

$X_2 \sim N(\mu_2, \sigma_2^2)$

$$p(X_1, X_2) = \text{NormPDF}(\mu_1, \sigma_1^2) \cdot \text{NormPDF}(\mu_2, \sigma_2^2)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \cdot e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_1} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \cdot e^{-\frac{1}{2}(x-\mu)^T \begin{bmatrix} \frac{1}{\sigma_1^2} & \\ & \frac{1}{\sigma_2^2} \end{bmatrix} (x-\mu)}$$

where $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ $\Sigma = \begin{bmatrix} \sigma_1^2 & \\ & \sigma_2^2 \end{bmatrix}$

$$\Rightarrow = \frac{1}{\sqrt{2\pi}\sigma_1} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}} \cdot (2\pi)^{\frac{1}{2}}} \cdot \frac{1}{\sigma_1} \cdot \frac{1}{\sigma_2} \cdot e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$$= \frac{1}{(2\pi)^{D/2}} \cdot \frac{1}{|\Sigma|^{1/2}}$$

Definition

Let X be a D-dim. random variable $X = [x_1, x_2, \ldots, x_D]^T_{D\times 1}$

$X \sim MVN(\mu, \Sigma)$

Parameters:

- "mean vector" $\mu \in \mathbb{R}^D$
- "covariance" Σ is $D \times D$

PDF:

$$MVN\ PDF(x; \mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right]$$

C = $\frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}}$

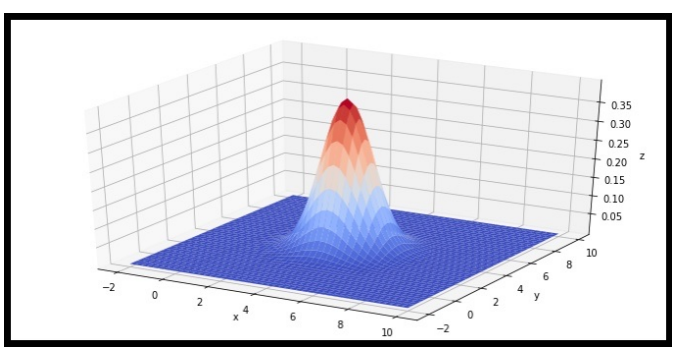
Mahalanobis distance

$f(x, \mu, \Sigma)$ = $\exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right]$

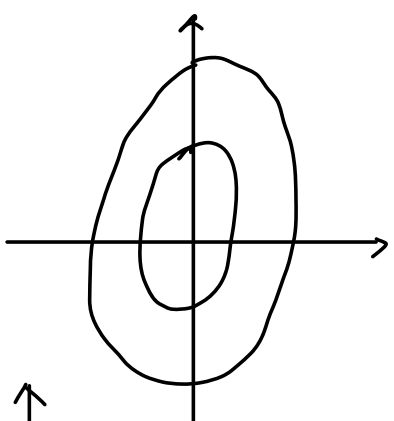
dimension check:

C: scaler $\times$ scaler $\in \mathbb{R}$

$f(x, \mu, \Sigma)$: $1\times D \times D \times D \times 1 = 1\times 1 \in \mathbb{R}$



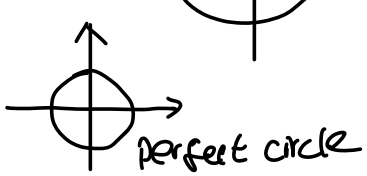
3d-plot



2d-plot

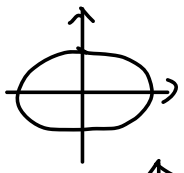
ellipses

$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

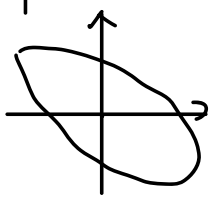


perfect circle

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



$\Sigma = \begin{bmatrix} 3 & -0.5 \\ -0.5 & 1 \end{bmatrix}$



Some properties:

$E(X) = \mu$, $Var(X) = \Sigma$. $\boxed{E(XX^T) = \Sigma + \mu\mu^T}$ $D \times D$

Covariance Properties: Why Contours are Elliptical

(Symmetric Positive Definite Matrix)

由于方差矩阵 Σ 是正定且对称的，对其进行特征值分解：

$\Sigma = Q\Lambda Q^{-1} = Q\Lambda Q^T$，其中 $QQ^T = Q^TQ = I$，$\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$

$$\Rightarrow \Sigma = (q_1\ q_2\ \cdots\ q_p)_{p\times p} \cdot \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}_{p\times p} \begin{bmatrix} q_1^T \\ q_2^T \\ \vdots \\ q_n^T \end{bmatrix}$$

$$= \begin{bmatrix} q_1\lambda_1 & q_2\lambda_2 & \cdots & q_n\lambda_n \end{bmatrix}_{1\times p} \begin{bmatrix} q_1^T \\ \vdots \\ q_n^T \end{bmatrix}_{p\times 1} \rightarrow 1\times 1$$

$$\Sigma = \sum_{i=1}^{p} q_i\lambda_i q_i^T$$

$$\Rightarrow \Sigma^{-1} = (Q\Lambda Q^T)^{-1}$$

$$= (Q^T)^{-1} \cdot \Lambda^{-1} \cdot Q^{-1} \quad \text{对正交矩阵 } Q^{-1} = Q^T$$

$$= Q \cdot \Lambda^{-1} \cdot Q^{-1} \quad \Lambda^{-1} = diag(\tfrac{1}{\lambda_i})$$

$$\Sigma^{-1} = \sum_{i=1}^{p} q_i \tfrac{1}{\lambda_i} q_i^T$$

$$(X-\mu)^T \Sigma^{-1} (X-\mu) = (X-\mu)^T \sum_{i=1}^{D} q_i \tfrac{1}{\lambda_i} q_i^T (X-\mu)$$

$$= \sum_{i=1}^{D} \frac{1}{\lambda_i} (X-\mu)^T q_i\ q_i^T (X-\mu) \quad *$$

Let $(X-\mu)^T q_i = y_i$. then $q_i^T (X-\mu) = y_i^T$

$$* \Rightarrow \sum_{i=1}^{D} \frac{1}{\lambda_i} y_i y_i^T = \boxed{\sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}}$$

Let $D = 2 \Rightarrow \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} = 1$

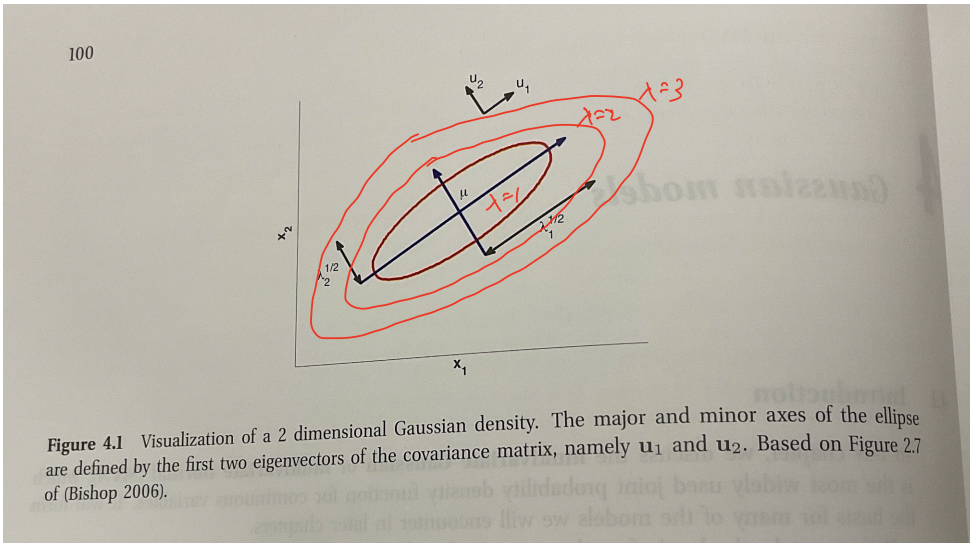


Figure 4.1 Visualization of a 2 dimensional Gaussian density. The major and minor axes of the ellipse are defined by the first two eigenvectors of the covariance matrix, namely $\mathbf{u}_1$ and $\mathbf{u}_2$. Based on Figure 2.7 of (Bishop 2006).

the eigenvectors determine the orientation of the ellipse, and the eigenvalues determine how elongated it is.

Margins of Gaussian are Gaussian

Suppose we have D-dimensional r.v. X

$X \sim MVN(\mu, \Sigma)$ $\quad \mu \in \mathbb{R}^D$. $\quad \Sigma \in \mathbb{R}^{D \times D}$

Consider any partition of indices

$\{1, 2, 3, \ldots, D-2, D-1, D\}$

first D^A dims belong to A → remaining dims belongs to B

then $\begin{bmatrix} X_A \\ X_B \end{bmatrix} = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix} \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}$, where $\Sigma_{AB} = \Sigma_{BA}^T$

We have the margin of A:

$p(X_A) = \int p(X_A, X_B)\, dB$

$= MVNormPDF(X_A | \mu_A, \Sigma_{AA})$

By symmetry:

$p(X_B) = \int p(X_A, X_B)\, dA$

$= MVNormPDF(X_B | \mu_B, \Sigma_{BB})$

Conditionals of a Joint Gaussian are Gaussian

Given joint distribution over X_A, X_B

$$\begin{bmatrix} X_A \\ X_B \end{bmatrix} = MVN\left(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{AB}^T & \Sigma_{BB} \end{bmatrix} \right)$$

Define $\Lambda = \Sigma^{-1}$

$$\Rightarrow \quad \Lambda = \begin{bmatrix} \Lambda_{AA} & \Lambda_{AB} \\ \Lambda_{AB}^T & \Lambda_{BB} \end{bmatrix}$$

Define the conditional density of X_A given $X_B = m_B$

$$P(X_A \mid X_B = m_B) = \frac{P(X_A, X_B = m_B)}{P(X_B = m_B)}$$

$$= \dots = MVNormPDF\left(X_A \mid \underbrace{\mu_A}_{A\times 1} - \underbrace{\Lambda_{AA}^{-1}}_{A\times A} \underbrace{\Lambda_{AB}}_{A\times B} \underbrace{(m_B - \mu_B)}_{B\times 1}, \underbrace{\Lambda_{AA}^{-1}}_{A\times A}\right)$$

Linear Gaussian model is Gaussian

Consider a model with two variables:

1) $x \sim MVN(\mu, \Lambda^{-1})$ $\quad x \in \mathbb{R}^S$

2) $y \mid x \sim MVN(\underset{T\times S}{A} \underset{S\times 1}{x} + \underset{T\times 1}{b}, \underset{T\times T}{L^{-1}})$ $\quad y \in \mathbb{R}^T$

What is the joint distribution? $P(x, y)$

$$P(x, y) = P(x) \cdot P(y \mid x)$$

$$\log P(x, y) = \log P(x) + \log P(y \mid x)$$

$$= \text{const} - \underbrace{\frac{1}{2}(x-\mu)^T \Lambda (x-\mu)}_{A} - \underbrace{\frac{1}{2}(y - Ax - b)^T L (y - Ax - b)}_{B}$$

$$A: \frac{1}{2}(x-\mu)^T \Lambda (x-\mu) = \frac{1}{2}(x^T \Lambda x - x^T \Lambda \mu - \mu^T \Lambda x + \underbrace{\mu^T \Lambda \mu}_{C})$$

$$= \frac{1}{2} x^T \Lambda x - \frac{1}{2} \cdot 2 (x^T \Lambda \mu)$$

$$B: \frac{1}{2}(y - Ax - b)^T L (y - Ax - b) = -\frac{1}{2} x^T A^T L A x - \frac{1}{2}(-2) y^T L A x -$$

$$\frac{1}{2} y^T L y - \frac{1}{2}(2) x^T A^T L b - \frac{1}{2}(-2) y^T L b$$

order 2 order 1

$$\Rightarrow -\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^T \underbrace{\begin{bmatrix} & P & \end{bmatrix}\begin{pmatrix} x \\ y \end{pmatrix}}_{\text{order } 2} + \underbrace{\begin{pmatrix} x \\ y \end{pmatrix}^T \begin{bmatrix} p_m \end{bmatrix}}_{\text{order } 1}$$

$$P = \begin{bmatrix} \Lambda + A^T L A & -A^T L \\ -LA & L \end{bmatrix}$$

$$m = P^{-1}\begin{bmatrix} \Lambda\mu - A^T L b \\ Lb \end{bmatrix}$$

Lec 8

1. Probabilistic view of linear regression 3.1 - 3.3
2. ML estimation of weights and precision
3. Towards a full probabilistic model for regression
4. Posterior of Gaussian - Gaussian linear regression
5. MAP estimation of weights

Probabilistic view of linear regression

Goal: Given dataset $\{x_n, t_n\}_{n=1}^N$, with $x_n \in \mathbb{R}^D$ and $t_n \in \mathbb{R}$

Want to predict $t_\#$ given $x_\#$

Assume: "linear model", which means prediction function is linear function of input x_n

$$y(x_n, w) = w_0 + w_1 x_{n1} + w_2 x_{n2} + \cdots + w_n x_{nD}$$
$$= \sum_{d=0}^{D} w_n x_{nd} \quad (\text{define } x_{n0} = 1) \quad x_{nd} \in D^{n+1}$$
$$= w^T x_n$$

- often can define "smarter" features by transforming input x_n into another feature space via $\phi(x_n)$

  Define $\phi(x_n) = [1 \quad \phi_1(n) \quad \phi_2(n) \quad \cdots \quad \phi_{M-1}(n)]$ — M total entries

  Note: $\phi(x_n)$ can be nonlinear, e.g. Gaussian...

- "Featured" model for prediction:

$$y(x_n, w) = \sum_{m=1}^{M} w_m \phi_m(x_n) = w^T \phi(x_n)$$

$w = [w_0, \ldots, w_{M-1}]^T$ $M \times 1$

$\phi = [\phi_0, \ldots, \phi_{M-1}]^T$ $M \times 1$

Note: our prediction will not be perfect. Need to tolerate some noise. Let's define a probabilistic approach.

- Likelihood of observing output t_n given input x_n $\quad \xi \sim N(0, \sigma^2)$

$$p(t_n \mid x_n, w, \beta) = N(t_n \mid w^T \phi(x_n), \beta^{-1}) \qquad t_n = w^T \phi(x_n) + \xi$$

mean (↑ 1-d variable in $\mathbb{R}$), variance

$w^T \phi(x_n)$ is a scaler, linear transformation of Gaussian is Gaussian

$x \sim N(\mu, \sigma^2)$. $Y = ax + b \sim N(a\mu + b, a^2\sigma^2)$

here $x \sim N(0, \sigma^2)$. $N(b, a^2\sigma^2)$

$N(b, \frac{1}{\beta^{-1}})$

If we assume all N observations are i.i.d. from this distribution, MLE:

$$p(t_n \mid x_n, w, \beta) = \prod_{i=1}^{N} N(t_n \mid w^T \phi(x_n), \beta^{-1})$$

$$\log p(t_n \mid x_n, w, \beta) = \frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi) - \beta \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T \phi(x_n))^2$$

key idea: use feature ϕ to get flexible prediction. Assume Gaussian noise model $N(t \mid w^T \phi(x_n), \beta^{-1})$

ML estimation of weights and precision

$$L(w, \beta) = \frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi) - \beta \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T \phi(x_n))^2$$

$$\frac{\partial L(w, \beta)}{\partial w} = -\frac{\beta}{2} \sum_{n=1}^{N} (t_n - w^T \phi(x_n))^2$$

($M \times 1$)

$$= \beta \sum_{i=1}^{N} (t_n - w^T \phi(x_n)) \cdot \underbrace{\phi(x_n)^T}_{M \times 1} \overset{\text{set}}{=} 0$$

$$= \sum_{i=1}^{N} t_n \phi(x_n)^T - \sum_{i=1}^{N} w^T \underbrace{\phi(x_n) \phi(x_n)^T}_{M \times M} \overset{\text{set}}{=} 0$$

$$\Rightarrow \underbrace{\sum_{i=1}^{N} t_n \phi(x_n)^T}_{1 \times M} = \underbrace{\sum_{i=1}^{N} w^T \phi(x_n) \phi(x_n)^T}_{1 \times M \quad M \times M} \quad *$$

Let $\Phi = \begin{bmatrix} - \phi(x_1) - \\ - \phi(x_2) - \\ \vdots \\ - \phi(x_n) - \end{bmatrix}_{n \times m}$, $t = \begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix}_{n \times 1}$

$*$ $\Rightarrow$ $t \Phi = w^T (\Phi^T \Phi)$

take transpose: $t \Phi^T = \Phi^T \Phi w$ $\quad \Phi^T \Phi$ has to be invertible!

$$w = (\Phi^T \Phi)^{-1} \Phi^T t$$

$m \times n \quad n \times m \quad m \times m \quad m \times n \quad n \times 1 \rightarrow m \times 1$ ✓

也可以用 MLE 求使！

$$L(w) = \| wX - y \| = \sum_{i=1}^{n} \| wx_i - y_i \|_2$$

- Penalized MLE

$$\max_w \mathcal{L}(w, \beta) + \lambda \sum_{m=1}^{M} w_m^2 \quad \text{(Ridge Regression)}$$

$$\hookrightarrow \quad w^* = (\lambda I_M + \Phi^T \Phi)^{-1} \Phi^T t$$

this is always rank M and always invertible!

Towards a full probabilistic model for regression

All unknown parameters (weight vector $w \in \mathbb{R}^M$, precision β^{-1}) are treated probabilistically. For now, we'll assume β^{-1} is fixed known.

Equivalently, we need to define a joint model

$$p(t, w \mid X, \beta) \qquad \text{(posterior} \propto \text{likelihood} \times \text{prior)}$$

$$= p(t \mid w, X, \beta) \cdot p(w)$$

why? Given this joint, we can talk about posterior about parameters after seeing data:

$$p(w \mid \{x_n, t_n\}_{n=1}^N, \beta)$$

- can use the MAP estimate instead of ML
- can use samples from posterior to access uncertainty

We can also use posterior to make predictions about new data

use predictive posterior

$$p(t_* \mid x_*, \{x_n, t_n\}_{n=1}^N, \beta) = \int_w p(t_*, w \mid x_*, \{x_n, t_n\}_{n=1}^N, \beta)\, dw$$

$$= \int_w p(t_{w_*} \mid w, x_*, \beta) \cdot p(w \mid \{x_n, t_n\}_{n=1}^N, \beta)\, dw$$

Remark:

$$p(x = x_* \mid D) = \int p(x = x_*, \mu \mid D)\, d\mu$$

$$= \int p(x = x_* \mid \mu, D) \cdot p(\mu \mid D)\, d\mu$$

$$\text{i.i.d.} \quad = \int p(x = x_* \mid \mu) \cdot p(\mu \mid D)\, d\mu$$

key difference: Average overall w vectors, weighted by posterior density. Do not just commit 100% to one w vector.

Review Gaussian

1. Given

Two independent Gaussian r.v.s

$X \sim N(\mu_x, \sigma_x^2)$

$Y \sim N(\mu_y, \sigma_y^2)$

Their joint is Gaussian

$$\begin{bmatrix} X \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}\right)$$

2. Two linearly-dependent Gaussin r.v.s

$X \sim N(\mu_x, \beta_x^{-1})$

$Y \sim N(mx+b, \beta_y^{-1})$

Their joint is Gaussian

$$\begin{bmatrix} X \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \beta_x + \beta_y m & -\beta_y m \\ -\beta_y m & \beta_y \end{bmatrix}^{-1}\right)$$

3. A joint distribution over a partitioned vector

$$\begin{bmatrix} \text{——} & x & \text{——} \end{bmatrix}^T = \begin{bmatrix} - & x_A & - \end{bmatrix} \begin{bmatrix} - & x_B & - \end{bmatrix}^T$$

$$\begin{bmatrix} x_A \\ x_B \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{AB}^T & \Sigma_{BB} \end{bmatrix}\right)$$

Marginal is Gaussian

$p(x_A) = N(\mu_A, \Sigma_{AA})$

Conditional is Gaussian

$p(x_A | x_B) = N(\mu_{A|B}, \Lambda_{AA}^{-1})$

$\mu_{A|B} = \mu_A - \Lambda_{AA}^{-1} \Lambda_{AB} (x_B - \mu_B)$

4. Two linearly dependent Gaussians.

$x \sim N(\mu, \Lambda^{-1})$

$y|x \sim N(Ax+b, L^{-1})$

Posterior is Gaussian

$p(x|y) = N(\mu_{x|y}, \Sigma)$

Marginal and Conditional Gaussians

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \quad (2.113)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}) \quad (2.114)$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \quad (2.115)$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \quad (2.116)$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}. \quad (2.117)$$

Posterior of Gaussian-Gaussian linear regression.

Assume $\beta^{-1} > 0$ known

Prior: $p(w) = \mathcal{N}(m_0, S_0)$ $\quad m_0 \in \mathbb{R}^M$, S_0 is $M \times M$ covariance matrix

$p(x)$

Likelihood:

$$p(t \mid x, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid w^T \phi(x_n), \beta^{-1})$$

$p(y|x)$ | why? $t = w^T \phi(x) + \varepsilon \rightarrow \mathcal{N}(0, \beta^{-1})$

$$p(t \mid x, w, \beta^{-1}) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid w^T \phi(x_n), \beta^{-1})$$

$\Rightarrow$

Posterior:

$$p(w \mid t, X, \beta) = \mathcal{N}(m_N, S_N)$$

$p(x|y)$

where $m_N = S_N (S_0^{-1} m_0 + \beta \Phi^T t)$ $\quad (3.50 - 3.51)$

Gaussian × Gaussian → Gaussian

$$S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi$$

where: $\Phi = \begin{bmatrix} - \phi(x_1) - \\ - \phi(x_2) - \\ \vdots \\ - \phi(x_n) - \end{bmatrix}_{n \times m}$, $t = \begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix}_{n \times 1}$

What is the MAP estimator for linear regression?

$$w_{MAP} = \underset{w \in \mathbb{R}^M}{\operatorname{argmax}} \log p(w \mid \{x_n, t_n\}_{n=1}^N)$$ ← we've shown this is $\mathcal{N}(m_N, S_N)$

thus. $w_{MAP} = m_N$

$$= \boxed{(S_0^{-1} + \beta \Phi^T \Phi)^{-1} (S_0^{-1} m_0 + \beta \Phi^T t)} \quad *$$

Consider a prior that favors zero mean and diag covariance

$\Rightarrow m_0 = \vec{0}$, $S_0 = \begin{bmatrix} \alpha^{-1} & & \\ & \ddots & \\ & & \alpha^{-1} \end{bmatrix} = \alpha^{-1} I_m$. $S_0^{-1} = \alpha I$ $\quad **$

plug $**$ into $*$:

$$(\alpha I_m + \beta \Phi^T \Phi)^{-1} \beta \Phi^T t$$

$$= \frac{\beta}{\beta} \left(\frac{\alpha}{\beta} I_m + \Phi^T \Phi \right)^{-1} \Phi^T t$$

$$= \boxed{\left(\frac{\alpha}{\beta} I_m + \Phi^T \Phi \right)^{-1} \Phi^T t}$$

$w^* = (\lambda I_M + \Phi^T \Phi)^{-1} \Phi^T t$

Ridge Regression!

Equivalent to our sum-of-squares penalized linear regression

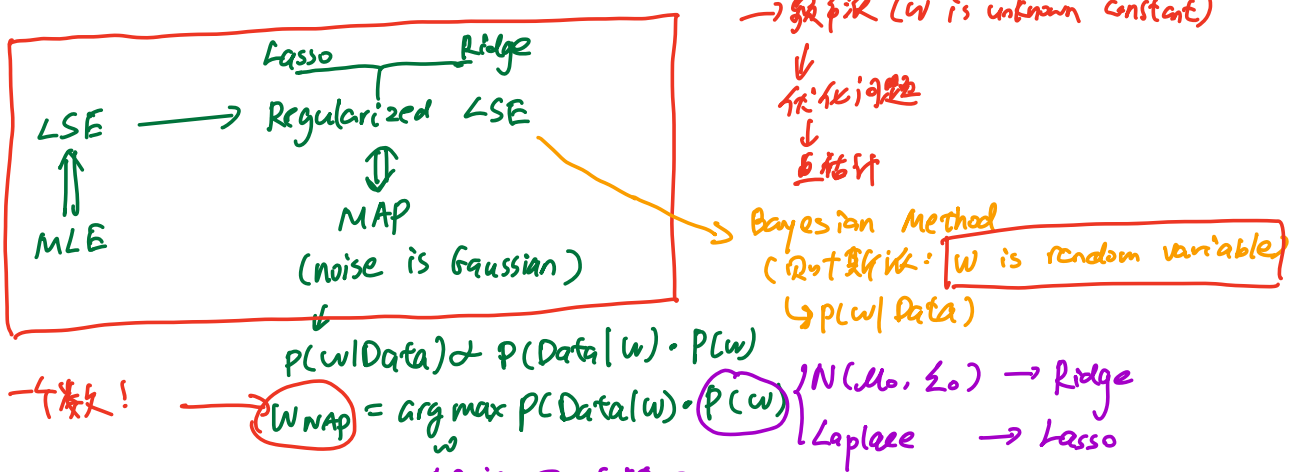
Conclusion:

Our MAP for Gaussian - Gaussian model is equal to ridge regression, under the constriant that

- prior is symmetric across w with precision λ
- known likelihood pre β

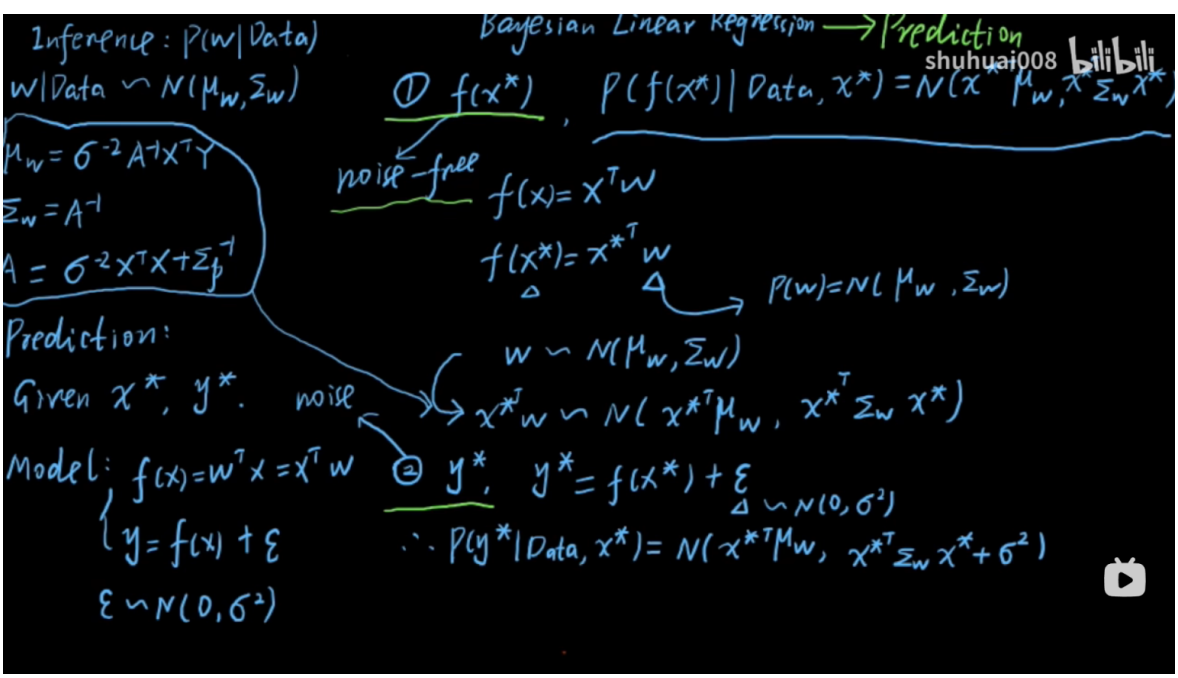
Supplement

1.

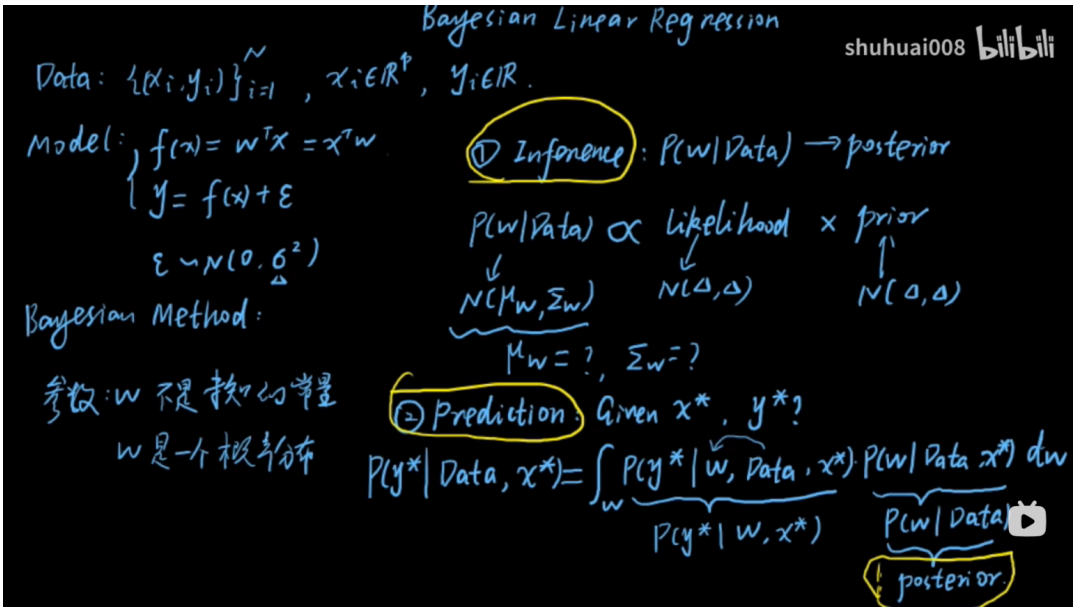


2. 算出来 $P(w|Data)$ 后，如何预测下一个值？

即 x^*, y^* ?



3. 总结.


Bayesian Linear Regression
shuhuai008 bilibili
Data: $\{(x_i, y_i)\}_{i=1}^{N}$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$.
Model: $f(x) = w^T x = x^T w$
$y = f(x) + \varepsilon$
$\varepsilon \sim N(0, \sigma^2)$
Bayesian Method:
参数: w 不是我们的常量
w 是一个概率分布
① Inference: $P(w|Data) \rightarrow$ posterior
$P(w|Data) \propto$ likelihood $\times$ prior
$N(\mu_w, \Sigma_w)$ $N(0, \sigma)$ $N(0, \Delta)$
$\mu_w = ?$, $\Sigma_w = ?$
② Prediction: Given x^*, y^*?
$P(y^*|Data, x^*) = \int_w P(y^*|w, Data, x^*) P(w|Data, x^*) dw$
$P(y^*|w, x^*)$
$P(w|Data)$
posterior

Lec 9

1. Posterior predictive 3.3.1 - 3.3.2
2. Evidence : A measure of model quality 3.5
3. Model selection 3.4
4. Hyperparameter estimation

Posterior predictive

Recall our model:

Prior: $p(w|\alpha) = \mathcal{N}(w|\, 0, \alpha^{-1}I)$ zero-mean isotropic Gaussian

Likelihood: $p(t|w,\beta) = \prod_{i=1}^{N} \mathcal{N}(t_n | w^T\phi(x_n), \beta^{-1})$

iid assumption, says if we saw a new observation at x_*, then

$$p(t_*, t | w, \beta) = p(t|w,\beta) \cdot p(t_*|w,\beta)$$

Then by linear Gaussian rules:

Joint distribution $p(t_{1:N}, w | \alpha, \beta)$ is Gaussian

$p(w,t) = \underbrace{p(w)}_{G} \cdot \underbrace{p(t|w)}_{G}$

Expanded Joint distribution $p(t_*, t_{1:N}, w | \alpha, \beta)$ is Gaussian *

Margin of * : $p(t_*, t_{1:N} | \alpha, \beta)$ is Gaussian. **

Condition of ** : $p(t_* | t_{1:N}, \alpha, \beta)$ is Gaussian

predictive posterior = p(new data | train data, hyperparameters)

Formulas For Predictive Posterior

$$p(t_* \mid t_{1:N}, \alpha, \beta) = \mathcal{N}\left(t_* \mid M_N^T \phi(X_*),\ \frac{1}{\beta} + \phi(X_*)^T S_N \phi(X_*)\right)$$

where
- M_N $M \times 1$ is posterior mean vector
- S_N $M \times M$ is posterior covariance vector
- $\phi(X_*)$ $M \times 1$ is feature vector at test point X_*
- β 1×1 is scaler likelihood precision

Useful Properties

- Average over many weights w, do not commit to a point estimate
$$p(t_* \mid t) = \int p(t_*, w \mid t)\, dw = \int p(t_* \mid w, t) \cdot p(w \mid t)\, dw$$
$$\overset{\downarrow \text{iid}}{=} \int p(t_* \mid w) \cdot p(w \mid t)\, dw$$
- Variance may change with location of prediction X_*. Will always be at least $\frac{1}{\beta}$. can be larger. $\frac{1}{\beta} + \underbrace{\phi(X_*)^T S_N \phi(X_*)}_{\geq 0}$
- Variance $\sigma_N^2(X_*)$ can not increase as more data seen.

Evidence: A measure of model quality

★ $w \in \mathbb{R}^M$
† all N observations. $t_n \in \mathbb{R}$

- $p(w)$ prior $p(w \mid \alpha, \beta)$
- $p(t \mid w)$ likelihood $p(t \mid w, \alpha, \beta)$
- $p(w \mid t)$ posterior $p(w \mid t, \alpha, \beta)$
- $p(t)$ evidence $p(t \mid \alpha, \beta)$

fixed hyperparameters
$\alpha > 0$
$\beta > 0$
free to add in!!! ★★

Intuition: Similar to posterior predictive except we weight by prior instead of posterior. We do it for the whole given dataset.

$$p(t \mid \alpha, \beta) = \int \underbrace{p(t \mid w, \beta)}_{\text{likelihood}} \cdot \underbrace{p(w \mid \alpha)}_{\text{prior}} \, dw$$

$p(y) = \int p(y \mid x) \cdot p(x) \, dx$ marginal probability

where: prior $p(w \mid \alpha) = \mathcal{N}(w \mid 0, \alpha^{-1} I)$

likelihood $p(t \mid w, \beta) = \mathcal{N}(t_n \mid w^T \phi(x_n), \beta^{-1})$

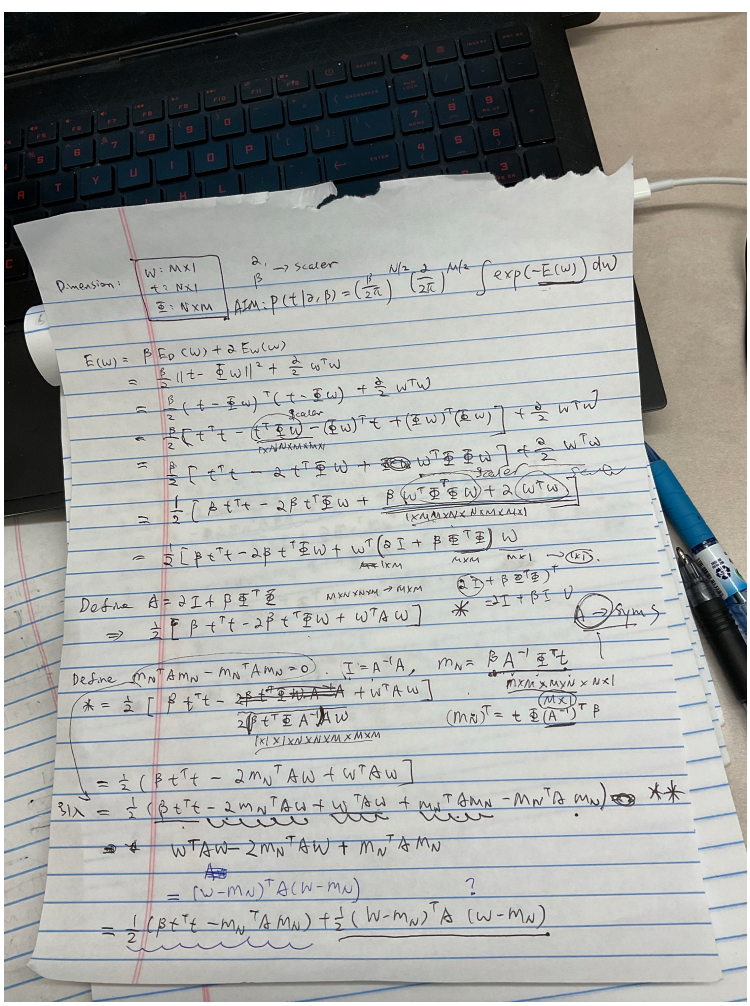
Predictive: $p(t_* \mid t_{1:N})$

$= p(t_* \mid t_{1:N}, x_*)$

$= \int_w \underbrace{p(t_* \mid w, x_*)}_{\text{likelihood}} \cdot \underbrace{p(w \mid \{x_n, t_n\}_{n=1}^N)}_{\text{Posterior}} \, dw$

经过推导

$$\ln p(\mathbf{t} \mid \alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(\boldsymbol{m}_N) - \frac{1}{2} \ln |\boldsymbol{A}| - \frac{N}{2} \ln(2\pi)$$



hint:

$p(x) = \mathcal{N}(\mu, \Sigma)$

指数部分：

$\exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))$

$= -\frac{1}{2}(x^T\Sigma^{-1} - \mu^T\Sigma^{-1})(x-\mu)$

$= -\frac{1}{2}(x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu - \mu^T\Sigma^{-1}x + \mu^T\Sigma^{-1}\mu)$

$= -\frac{1}{2}(x^T\Sigma^{-1}x - 2\mu^T\Sigma^{-1}x + C)$

$x^T\Sigma^{-1}x$ → 二次项　　$\mu^T\Sigma^{-1}x$ → 一次项

$A = \alpha I + \beta \Phi^T \Phi$

Define: $m_N^T A m_N - m_N^T A m_N = 0$ $\quad I = A^{-1}A$

$m_N = \beta A^{-1} \Phi^T t$

接上:

$\frac{1}{2}(\beta t^T t - m_N^T A m_N) + \frac{1}{2}(w - m_N)^T A (w - m_N)$

$\frac{1}{2}(\beta t^T t - m_N^T A m_N)$ 目标 $\frac{\beta}{2}\|t - \Phi m_N\|^2 + \frac{\alpha}{2} m_N^T m_N$ 没有 A

$= \frac{1}{2}(\beta t^T t - m_N^T A m_N + m_N^T A m_N - m_N^T A m_N)$

$= \frac{1}{2}(\beta t^T t - 2 m_N^T A m_N + m_N^T A m_N)$

$= \frac{1}{2}(\beta t^T t - 2 m_N^T A \beta A^{-1} \Phi^T t + m_N^T (\alpha I + \beta \Phi^T \Phi) m_N^T)$

$= \frac{1}{2}(\beta t^T t - 2 m_N^T \Phi^T t \beta + \alpha m_N^T m_N + \beta m_N^T \Phi^T \Phi m_N)$

$= \frac{1}{2}(\beta t^T t - 2 m_N^T \Phi^T t \beta + \beta m_N^T \Phi^T \Phi m_N) - \frac{\alpha}{2} m_N^T m_N$ 待

$= \frac{\beta}{2}(t^T t - 2 m_N^T \Phi^T t + m_N^T \Phi^T \Phi m_N)$

$t^T t - 2 t m_N^T \Phi^T + m_N^T \Phi^T \Phi m_N$

$a^2 \quad 2ab \quad b^2$

where $a = t \quad b = m_N^T \Phi^T$

$= \frac{\beta}{2}\|t - \Phi m_N\|_2^2 + \frac{\alpha}{2} m_N^T m_N)$ proceed

$E(w) = \frac{\beta}{2}\|t - \Phi m_N\|_2 + \frac{\alpha}{2} m_N^T m_N$ ← define as

目标: $\int \exp(-E(w))\, dw$

$= \int \exp(-\frac{\beta}{2}\|t - \Phi m_N\|_2 + \frac{\alpha}{2} m_N^T m_N)$

Given $S_n^{-1} = \alpha I + \beta \Phi^T \Phi$, means $A = S_n^{-1}$

$\Rightarrow m_N = \beta S_n \Phi t$ ⇐ 后验分布的均值

假设完后面有部分待证.

Model Selection
(Fixed validation/cross validation / Evidence)

use the evidence for selection

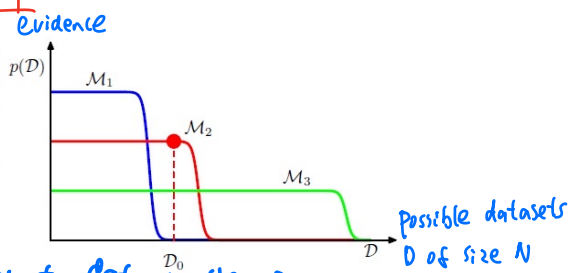


Figure 3.13 Schematic illustration of the distribution of data sets for three models of different complexity, in which $\mathcal{M}_1$ is the simplest and $\mathcal{M}_3$ is the most complex. Note that the distributions are normalized. In this example, for the particular observed data set $\mathcal{D}_0$, the model $\mathcal{M}_2$ with intermediate complexity has the largest evidence.

have unit area ←

- simplier models give higher density to dataset they favor.
- Using evidence will avoid overfitting

Usual Procedure: Pick single best model M^* for L possible. Use that to make predictions.

Bayesian Procedure: Average all models, weighted by posterior:

$$p(t_* \mid t) = \sum_{\ell=1}^{L} p(t_* \mid t, M_\ell) \cdot p(M_\ell \mid t)$$

$p(t_* \mid t, M_\ell)$: predictive posterior for model ℓ; $p(M_\ell \mid t)$: model posterior

Comparison of Methods for Hyperparameter Selection

	Fixed valid. set (fraction f)	K-fold cross-validation	Bayesian evidence	
Fraction data used for training run	(1.0 – f)	(K-1) / K	100% 1.0	**Higher is better** Better use of training data
Total runs/ examples seen for training	1 run (1 – f) N	K runs (K-1) * N	1 run N	**Lower is better** Faster training
Total runs/ examples seen for evaluation of fitness	1 run fN	K runs N	1 run N	**Lower is better** Faster evaluation
Fitness function	Heldout likelihood	Heldout likelihood	Evidence	
complexity	low	low	HARD	

Hyperparameter estimation

Want to know good values for α, β given dataset.

$$\hat{\alpha}, \hat{\beta} = \underset{\substack{\alpha>0 \\ \beta>0}}{\arg\max} \log p(t \mid \alpha, \beta)$$

This is "empirical Bayesian" point estimation

Can be solved via:

1. enumeration via grid search
2. gradient descent
3. coordinate descent (will see in EM algorithm)
4. analytical estimates

Lec 10 Probabilistic Generalized linear models

1. Discriminative & Generative model
2. Generalized Linear models
3. Sigmoid function
4. Probabilistic Logistic Regression: ML & MAP strategy
5. 2nd-order gradient methods for linear & logistic Regression

Discriminative & Generative model

Consider supervised learning task:
Given training data $\{x_n, y_n\}_{n=1}^N$. Learn to predict $p(y_* | x_*)$

Two approaches to probabilistic modeling.

Discriminative (判别模型)

Model:
$p(y|x) = \prod_{n=1}^{N} p(y_n | x_n)$
(x is treated as fixed / known)

Parameter:
w generates model given feature

prediction: $p(y_* | x_*, w)$
(directly use likelihood)

Training: $\max_w \prod_n p(y_n | x_n, w)$

Pros: - simpler, fewer parameters
- directly solve supervised task

Con: - cannot predict if x has missing values.

Generative (生成模型)

Model:
$p(x,y) = \prod_{i=1}^{N} p_\theta(x_n | y_n) \cdot p_\pi(y_n)$
(both x & y are r.v.)

Parameters:
θ: generate x given y
π: generate y

Prediction via Bayes rule:

$$p(y_* | x_*) = \frac{p(x_* | y_*) \cdot p(y_*)}{p(x_*)}$$

$$= \frac{p(x_* | y_*) \cdot p(y_*)}{\sum p(x_* | y') \cdot p(y')}$$

Training: $\arg\max_{\theta, \pi} \log p_\theta(x_n | y_n) + \log p_\pi(y_n)$

Pro: - can predict if x / w is missing value
- can sample new x from $p(x|y)$

Cons: - more complex
- more parameters

Generalized Linear models

Extend linear regression to tasks with constrained output space

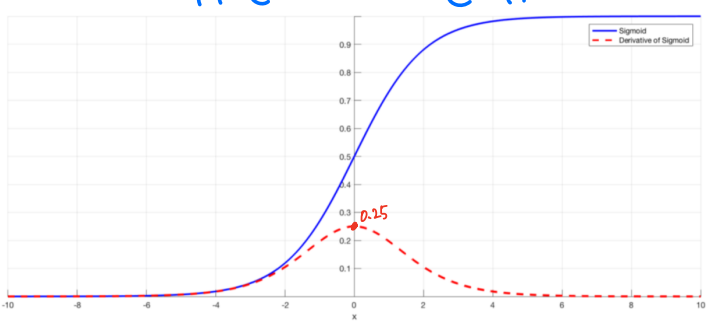
Training Data: $\{x_n, t_n\}_{n=1}^{N}$

Model: $p(t_{1:N} \mid x_{1:N}) = \prod_{n=1}^{N} Q(t_n \mid f(w^T \phi(x_n)))$

Output Space $\mathcal{Y}$	name	distribution Q	link function f
real $(-\infty, +\infty)$	linear regression	Normal	$f(w,x) = w^T \phi(x)$
integers $\{0, 1, 2, \cdots\}$	Poisson regression	Poisson	$f(w,x) = e^{w^T \phi(x)}$
binary $\{0, 1\}$	Logistic regression	Bernoulli	$f(w,x) = \sigma(w^T \phi(x))$ (logistic sigmoid)
	Probit regression	Bernoulli	$f(w,x) = \Phi(w^T \phi(x))$ (Φ: Normal CDF)
C-ary classification (1 0 0 ⋯ 0, 0 1 0 ⋯ 0, ⋮, 0 0 0 ⋯ 1)	Multi-class logistic regression	Categorical	$f(w,x) = \text{Softmax}(W^T \phi(x))$ (W: M×C matrix, vector in $\mathbb{R}^C$)
C levels of known ranking (e.g. ratings 0-5 stars; e.g. cold, warm, hot)	Ordinal regression	Ordinal	look up if interested!
positive reals $(0, +\infty)$	Exponential regression	Exponential	

Sigmoid Function

$$\sigma(r) = \frac{1}{1 + e^{-r}} = \frac{e^r}{e^r + 1}$$



properties:

- squashes real line to probability interval $(0,1)$
- monotonic: $r_B > r_A \Rightarrow \sigma(r_B) > \sigma(r_A)$
- Invertible: reals $(-\infty, \infty)$ $\xrightarrow{\text{Sigmoid}}$ probability $(0,1)$ $\xrightarrow{\text{logit}}$ reals; logit: $\ln \frac{p}{1-p}$
- $1 - \sigma(r) = \sigma(-r)$

- $\sigma'(r) = \sigma(r)(1-\sigma(r)) \rightarrow [0, 0.25]$

Probabilistic Logistic Regression: ML & MAP strategy

Prior: $p(w) = N(w|m_0, S_0)$

$= N(w|0, \alpha^{-1} I_M)$ usually

Likelihood: $p(t_n|w) = \text{BernPMF}(t_n|\sigma(w^T\phi(x_n)))$

Point Estimation Strategies:

MLE: $\arg\max_{w \in R^M} \sum_{i=1}^{N} \log p(t_n|w)$

MAP: $\arg\max_{w \in R^M} \sum_{i=1}^{N} \log p(t_n|w) + \log p(w)$

Next time: Posterior Estimation $p(w|t_{1:N})$

strategies for optimization.

$\underline{\arg\min}_{w \in R^M} a(w)$ always argmin - argmax 加负号变成 argmin

Strategies for Optimization

$\arg\min_{w \in \mathbb{R}^M} a(w)$

convention: always minimize

	ML/MAP for Linear Regression	ML/MAP for Logistic Regr.
Analytical methods		
- Set up $\nabla_w a = 0$	rarely works	
- Manipulate to find $w^* = \ldots$ closed form expression	but fast+exact once implemented ✓	✗ No closed form solution exists.
Gradient methods step size $\varepsilon_t > 0$, gradient $g \in \mathbb{R}^M$, Hessian $H \in \mathbb{R}^{M \times M}$ sym. p.d.		
- 1st order gradient descent $w_{t+1} \leftarrow w_t - \varepsilon_t g(w_t)$	Many iterations, each cheap: $O(M)$ ✓	✓
- 2nd order gradient descent $w_{t+1} \leftarrow w_t - \varepsilon_t H(w_t)^{-1} g(w_t)$	Few iters, each expensive: $O(M^3)$ inverting matrix ✓	✓
Other methods		
- grid search	✓	✓
- random search	✓	✓
- Nelder-Mead	✓	✓
- ... many others possible	usually very inefficient in high dimensions ($M > 3$) ✓	✓

2nd-order gradient methods for linear & logistic Regression

ML/MAP estimation: Gradients + Hessians

For Logistic Regression: $\mathcal{L}(w) = -\sum_{n=1}^{N} \log \text{BernPMF}(t_n \mid \sigma(w^T \phi(x_n))) - \log \text{NormPDF}(w \mid 0, \alpha^{-1} I_M)$

	gradient $\nabla_w \mathcal{L}$ shape: $(M, 1)$	Hessian $\nabla_w \nabla_w \mathcal{L}$ shape: (M, M)
Linear Regr.	$\beta \Phi^T (\Phi w - t) + \alpha w$ (Φw: predicted, t: true label)	$\beta \Phi^T \Phi + \alpha I_M$
Logistic Regr.	$\Phi^T (\sigma(\Phi w) - t) + \alpha w$ ($\sigma(\Phi w)$: prediction (probability), t: true binary label)	$\Phi^T R(w) \Phi + \alpha I_M$

where $R(w) = \begin{bmatrix} r_1(w) & & \\ & \ddots & \\ & & r_N(w) \end{bmatrix}$ ($N \times N$ diagonal matrix; all diagonal values ≥ 0, off-diagonal values $= 0$)

$r_n(w) = \sigma(w^T \phi(x_n)) \sigma(-w^T \phi(x_n))$ always > 0 since $\sigma(a) > 0$ for all a

$\mathcal{L}(w)$ — Unique minima w^* exists

Question:

Unique minima w^* exists when objective is convex

- Is Linear Regression MAP convex?

A: Yes. $\beta \underbrace{\Phi^T \Phi}_{\text{P.S.D}} + \underbrace{\alpha I_M}_{\text{P.S.D if } \alpha > 0}$.

- Is LR MAP convex?

A: Yes. $\Phi^T R(w) \Phi + \alpha I_M$

$R(w) \rightarrow$ positive diagonal

$\Phi^T R(w) \Phi$ is still positive

2nd order gradient methods

solve an optimization problem:

$$\underset{w \in \mathbb{R}^M}{\operatorname{argmin}}\ L(w)$$

Let $g(w) = \nabla_w L$ $\quad (M,1)$

$H(w) = \nabla_w \nabla_w L$ $\quad (M,M)$

1st. order GD:

$$w^{new} = w^{old} - \varepsilon\, g(w^{old})$$

2nd order GD:

$$w^{new} = w^{old} - \varepsilon\, H(w^{old})^{-1} g(w^{old})$$

Linear Regression: 2nd order GD

Using formulas from previous page:

$$w^{new} = w^{old} - \varepsilon \left[\beta \underline{\Phi}^T \underline{\Phi} + \alpha I_M\right]^{-1} \left(\beta \underline{\Phi}^T \underline{\Phi} w^{old} - \beta \underline{\Phi}^T t + \alpha w^{old}\right)$$

$$= w^{old} - \varepsilon \left(\beta \underline{\Phi}^T \underline{\Phi} + \alpha I_M\right)^{-1} \left[\left(\beta \underline{\Phi}^T \underline{\Phi} + \alpha I_M\right) w^{old} - \beta \underline{\Phi}^T t\right]$$

$$= w^{old} - \varepsilon\, w^{old} + \varepsilon \left(\beta \underline{\Phi}^T \underline{\Phi} + \alpha I_M\right)^{-1} \beta \underline{\Phi}^T t$$

If step size $\varepsilon = 1$, we get:

$$w^{new} = \left(\underline{\Phi}^T \underline{\Phi} + \frac{\alpha}{\beta} I_M\right)^{-1} \underline{\Phi}^T t$$

optimal MAP estimate in one step!

Anytime loss is quadratic, 2nd order gradient update with step size $\varepsilon = 1$ will find global minima in one step.

Logistic Regression w/ 2nd order GD

Still gold standard for ML/MAP estimation, but requires many iterations (unlike linear regression ML estimation using 2nd order methods).

Update looks like this (assuming ML estimation. MAP is similar.)

$$w^{new} \leftarrow w^{old} - \varepsilon \left(\underline{\Phi}^T R(w^{old}) \underline{\Phi} \right)^{-1} \underline{\Phi}^T \underbrace{\left(\sigma(\underline{\Phi} w^{old}) \right)}_{\hat{y}} - t \Big)$$

Assume $\varepsilon = 1$ and rearranging terms

$$\leftarrow \left(\underline{\Phi}^T R \underline{\Phi} \right)^{-1} \left(\underline{\Phi}^T R \underline{\Phi} w^{old} - \underline{\Phi}^T \hat{y} + \underline{\Phi}^T t \right)$$

$$\leftarrow \left(\underline{\Phi}^T R \underline{\Phi} \right)^{-1} \left(\underline{\Phi}^T R z \right)$$

where $z \in \mathbb{R}^N$

$z = \underline{\Phi} w^{old} - R^{-1}(\hat{y} - t)$

Interpret as "least squares" with per-example "weights" r, $R = \text{diag}(r)$ and outcomes z, both of which depend on previous value of w.

Thus, the name "Iteratively Re-weighted Least Squares" (IRLS)

Often used for Generalized Linear Models to do ML estimation of weights $w \in \mathbb{R}^M$

while not converged:
- update R given w
- update z given w
- update $w \leftarrow (\underline{\Phi}^T R \underline{\Phi})^{-1} \underline{\Phi}^T R z$

Supplement. 詳細推導 (Bishop 207 - 208)

$$W_{new} = W^{old} - H^{-1}\nabla E(W)$$

$E(W)$: the sum-of-square error function

→ Linear / Logistic Regression 都可以適用二階導來求解.

Let $\phi(X_n) = \phi_n$ ← Loss Function

Linear Regression

$$E(W) = \frac{1}{2}\sum_{n=1}^{N}(t_n - W^T\phi_n)^2$$

$$\nabla E(W) = \sum_{n=1}^{n}(t_n - W^T\phi_n)\cdot(-\phi_n)$$

$$= \sum_{n=1}^{n}(W^T\phi_n - t_n)\cdot(\phi_n)$$

$$= W^T\sum_{n=1}^{n}\phi_n\phi_n^T - \sum_{n=1}^{n}t_n\phi_n$$

$$\nabla E(W) = \Phi^T\Phi W - \Phi^T t$$

$$\Phi = \begin{bmatrix} -\varphi_1- \\ -\varphi_2- \\ \vdots \\ -\varphi_n- \end{bmatrix}_{N\times M} \qquad W = \begin{bmatrix} \\ \\ \end{bmatrix}_{M\times 1}$$

$$H = \nabla\nabla E(W) = \Phi^T\Phi$$

$$W_{new} = W^{old} - H^{-1}\nabla E(W)$$

$$= W^{old} - (\Phi^T\Phi)^{-1}(\Phi^T\Phi W^{old} - \Phi^T t)$$

$$= W^{old} - [W^{old} - (\Phi^T\Phi)^{-1}\Phi^T t)$$

$$= \boxed{(\Phi^T\Phi)^{-1}\Phi^T t}$$

Logistic Regression

$$W^{new} = W^{old} - H^{-1}\nabla E(W)$$

Bernoulli $t_n \in \{0,1\}$

$$p(t|W) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}$$

$$E(W) = -\ln p(t|W) = -\sum_{i=1}^{n}(t_n \ln y_n + (1-t_n)\ln(1-y_n)) \quad *$$

where $y_n = \sigma(W^T\phi_n)$

$$* = -\sum_{i=1}^{n}(\underbrace{t_n \ln \sigma(W^T\phi_n)}_{①} + \underbrace{(1-t_n)\ln(1-\sigma(W^T\phi_n))}_{②} \quad **$$

Given $\sigma'(x) = \sigma(1-\sigma)$

① $t_n \ln \sigma(w^T\phi_n)'$

$= t_n \cdot \frac{1}{\sigma(w^T\phi_n)} \cdot \sigma(w^T\phi_n) \cdot (1 - \sigma(w^T\phi_n)) \cdot \phi_n$

$= \sum_{i=1}^{n} t_n \phi_n [1 - \sigma(w^T\phi_n)) = \sum_{i=1}^{n} t_n\phi_n - t_n\phi_n \sigma(w^T\phi_n)$ ①

② $(1-t_n) \ln(1-\sigma(w^T\phi_n))'$

$= (1-t_n) \cdot \frac{1}{1-\sigma(w^T\phi_n)} \cdot -(\sigma(w^T\phi_n) \cdot (1-\sigma(w^T\phi_n)) \cdot \phi_n)$

$= \sum_{i=1}^{n} -(1-t_n) \cdot \phi_n \cdot \sigma(w^T\phi_n)$

$= \sum_{i=1}^{n} (\phi_n t_n - \phi_n) \cdot \sigma(w^T\phi_n)$

$= \sum_{i=1}^{n} \phi_n t_n \sigma(w^T\phi_n) - \phi_n \sigma(w^T\phi_n)$ ②

If ②: $- \sum_{i=1}^{n} (t_n \phi_n - \phi_n \sigma(w^T\phi_n))$ ($\sigma(w^T\phi_n) = y_n$)

$\nabla E(w) = \sum_{n=1}^{n} (y_n - t_n) \phi_n$

$= \sum_{n=1}^{n} y_n \phi_n - \sum_{i=1}^{n} \phi_n t_n$

$\boxed{\nabla E(w) = \Phi^T (y - t)}$

$\nabla\nabla E(w) = \sum_{n=1}^{n} (\phi_n \sigma(w^T\phi_n) - t_n\phi_n)'$

$= \sum_{n=1}^{n} (\phi_n \sigma(w^T\phi_n)(1 - \sigma(w^T\phi_n)) \cdot \phi_n)$

$= \sum_{n=1}^{n} (\phi_n^T y(1-y) \cdot \phi_n)$

$= \sum_{n=1}^{n} (\phi_n \phi_n^T) y(1-y) = \Phi^T R \Phi$

$$\nabla\nabla E(w) = \Phi^T R \Phi \quad \text{, where } R = \sum_n y_n(1-y_n)$$

$$\begin{aligned} w_{new} &= w_{old} - H^{-1}\nabla E(w) \\ &= w_{old} - (\Phi^T R \Phi)^{-1}(\Phi^T(y-t)) \\ &= w_{old} - [\Phi^T R \Phi)^{-1}(\Phi^T y - \Phi^T t) \end{aligned}$$

Update looks like this (assuming ML estimation. MAP is similar.)

$$w^{new} \leftarrow w^{old} - \varepsilon\left(\Phi^T R(w^{old})\Phi\right)^{-1}\Phi^T\left(\underbrace{\sigma(\Phi w^{old})}_{\hat{y}} - t\right)$$

Assume $\varepsilon = 1$ and rearranging terms

$$\leftarrow \left(\Phi^T R \Phi\right)^{-1}\left(\Phi^T R \Phi w^{old} - \Phi^T \hat{y} + \Phi^T t\right)$$

$$\leftarrow \left(\Phi^T R \Phi\right)^{-1}\left(\Phi^T R z\right) \quad \text{where } z \in \mathbb{R}^N$$

$$z = \Phi w^{old} - R^{-1}(\hat{y} - t)$$

Interpret as "least squares" with per-example "weights" r, $R = \text{diag}(r)$ and outcomes z, both of which depend on previous value of w.

Thus, the name "Iteratively Re-weighted Least Squares" (IRLS)

Often used for Generalized Linear Models to do ML estimation of weights $w \in \mathbb{R}^M$

Lec 11 Bayesian Logistic Regression 4.4 . 4.5

1. Overview of posterior & predictive
2. Laplace approximation in 1-dim and M-dim
3. Laplace approximation for Logistic regression
4. Posterior predictive approximation.

Overview of posterior & predictive

Model

Prior on weights $w \in \mathbb{R}^M$

$p(w) = N(w \mid m_0, S_0)$

Likelihood of "output" $t_n \in \{0, 1\}$

$p(t \mid w) = \prod_{n=1}^{N} \text{Bern}(t_n \mid \sigma(w^T \phi(x_n)))$ (iid)

Goal: To estimate the Posterior and Predictive

Posterior: $p(w \mid t_{1:N})$ No closed form!

Not a Gaussian (Not Gau x Gau again!)

Predictive: $p(t_* \mid t_{1:N})$

$= p(t_* \mid t_{1:N}, x_*)$

$= \int_w p(t_* \mid w, x_*) \cdot p(w \mid \{x_n, t_n\}_{n=1}^{N}) \, dw$

(likelihood · Posterior)

- Must be a Bernoulli
- But no close form for its parameter

Laplace approximation in 1-dim and M-dim

Given: a random variable $w \in \mathbb{R}$

$$p(w) = \frac{1}{Z} f(w)$$

Here, $f(w) > 0$ is known and differentiable

but computing $\int_w f(w)\, dw$ is hard

Q: How can we estimate the distribution $p(w)$?

Intuition: 用一个高斯的分布 $q(w)$ 来近似的分布 $p(w)$

Detail:

- pick mean to match the mode of $p(w)$

$$m = \arg\max_{w \in \mathbb{R}} p(w) = \arg\max_{w} f(w)$$

Can use Gradient methods to solve this numerically

- pick precision to perform best possible 2nd order Taylor approximation to $p(w)$ at the mode $w = m$

$$\beta = \frac{\partial}{\partial w}\frac{\partial}{\partial w}\left[-\log f(w)\right]\Big|_{w=m}$$

Advantages: Gives an approx distribution we can reason about.

Second derivatives are often tractable!

Limitations: bad if $p(w)$ is multimodal

bad if $p(w)$ { has heavy tails / not symmetric about mode.

Derivation of Taylor approx to density $p(w)$ at $m = \arg\max_w p(w)$

$$\log p(w) = \log f(w) + \text{const}$$

$$\triangleq \ell(w) + \text{const}$$

define $\ell(w) = \log f(w)$

$$= \ell(m) + \ell'(m)(w-m) + \frac{\ell''}{2}(w-m)^2 + C$$

2nd order Taylor approx to func ℓ at $w = m$

m is mode, $\ell'(m) = 0$

$$= -\frac{1}{2}\left[-\ell''(m)\right](w-m)^2 + C$$

$$= -\frac{1}{2}\beta(w-m)^2$$

where $\beta = \frac{1}{\sigma^2} = -\ell''(m)$ (precision)

mean: $m = \arg\max_m \ell(w)$

Laplace Approx in M-Dim

Idea: Approx: $q(w) = \text{MVNorm}(m, \Lambda^{-1})$ — m: mean, Λ: precision matrix

- Set m to match the mode of f

$$m = \arg\max_{w \in \mathbb{R}^m} \ell(w)$$

- Set Λ to negative Hessian at mode

$$\Lambda = \nabla_w \nabla_w [-\ell(w)] \big|_{w=m}$$

Summary: 用 Laplace approximation 近似一个分布，先寻找一个众数 z_0（驻点），然后计算那个众数位置上的 Hessian matrix.

- 根据中心极限定理，模型的后验概率分布会随着观测数据的增加而越来越近似于 Gaussian Distribution，因此可以在数据点较多的情况下，Laplace approximation 更有用！

Laplace approximation for Logistic regression

True posterior intractable, but known to constant via Bayes.

$$p(w \mid t_{1:N}) = \frac{p(t_{1:N} \mid w) \cdot p(w)}{p(t_{1:N})}$$

$$\Rightarrow \log p(w \mid t_{1:N}) = \log p(t_{1:N} \mid w) + \log p(w) - \underbrace{\log p(t_{1:N})}_{C}$$

$$= \underbrace{\log \text{MVNormPDF}(w \mid m_0, S_0) + \sum_{n=1}^{N} \text{Bern PMF}(t_n \mid \sigma(w^T \phi(x_n)))}_{\mathcal{L}(w)}$$

Apply Laplace approx:

$$p(w \mid \{x_n, y_n\}_1^N) \approx N(w \mid m_{MAP}, S)$$

where $m_{MAP} = \arg\max_w \mathcal{L}(w)$

$$S^{-1} = \nabla_w \nabla_w [-\mathcal{L}(w)] \big|_{w = m_{MAP}}$$

- Using formula for the hessian of MAP objective, we know:

$$S^{-1} = S_0^{-1} + \Phi^T R(m_{MAP}) \Phi$$

$$= \alpha I_m + \Phi^T R(m_{MAP}) \Phi$$

Recall that $\Phi^T R \Phi = \sum_{n=1}^{N} r(m_{MAP}, x_n)\, \phi(x_n)\, \phi(x_n)^T$

Key point: Laplace parameters m, S^{-1} possible to compute!

Predictive posteriors for Bayesian Logistic Regression

Ideal (intractable) posterior predictive:

$$p(t_* \mid t_{1:N}) = \int_{w \in \mathbb{R}^M} p(t_* \mid w)\, p(w \mid t_{1:N})\, dw$$

Laplace approximation

$p(w \mid t_{1:N}) \approx \mathcal{N}(m_{MAP}, S)$

$$\approx \int_{w \in \mathbb{R}^M} p(t_* \mid w)\, \mathcal{N}(w \mid m_{MAP}, S)\, dw$$

Still a tough integral over M dimensions

If M was 1 or 2, could use numerical strategies like trapezoid approx.

Option 1: Monte Carlo

Easy but need many samples L

$$p(t_* \mid t_{1:N}) = \mathop{\mathbb{E}}_{p(w \mid t_{1:N})} \left[p(t_* \mid w) \right]$$

Average of L samples from approx. posterior

$$\approx \frac{1}{L} \sum_{l=1}^{L} p(t_* \mid w^l)$$

with $w^l \sim \mathcal{N}(m_{MAP}, S)$

$$\int [t_* \mid w, t_{1:N})\, dw$$

$$= \int p(w \mid t_{1:N}) \cdot p(t_* \mid w)\, dw$$

$$= \mathop{\mathbb{E}}_{p(w \mid t_{1:N})} \left[p(t_* \mid w) \right]$$

Option 2: Probit approx.

Closed-form, hard to port to other models

Likelihood $p(t_* \mid w) = \begin{cases} \sigma(w^T \phi_*) & \text{if } t_* = 1 \\ \sigma(-w^T \phi_*) & \text{if } t_* = 0 \end{cases}$

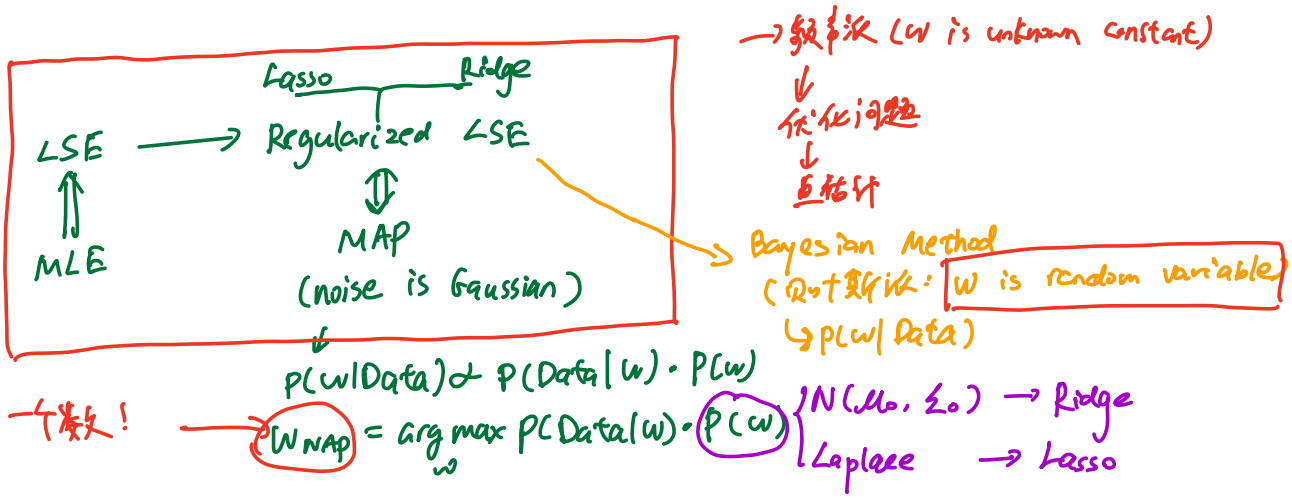
See Bishop Fig 4.9

$\sigma(a) \approx \text{NormCDF}(\sqrt{\frac{\pi}{8}}\, a)$

$$\approx \begin{cases} \text{NormCDF}(\sqrt{\frac{\pi}{8}}\, w^T \phi_*) & \text{if } t_* = 1 \\ 1 - \text{above} & t_* = 0 \end{cases}$$

Makes integral tractable when combined w/ Laplace!

(4.155): $p(t_* = 1 \mid t_{1:N}) = \sigma\left(m_{MAP}^T \phi_* \cdot \frac{1}{\sqrt{1 + \frac{\pi}{8} \phi_*^T S \phi_*}} \right)$


Lasso
Ridge
LSE
Regularized LSE
MLE
MAP
(noise is Gaussian)
→ 频率派 (w is unknown constant)
优化问题
点估计
Bayesian Method
(贝叶斯派: w is random variable)
↳ p(w|Data)
p(w|Data) ∝ p(Data|w) · p(w)
一个数!
w_MAP = arg max p(Data|w) · p(w)
N(μ0, Σ0) → Ridge
Laplace → Lasso