SPR Day 8

Bayesian Linear Regression

Concepts: Review Gaussian special properties

A	B
from conditional	to joint
from joint	to conditional
from joint	to marginal
from prior + lik	to posterior

if A is Gaussian, then B is too

Posterior for Gaussian linear regression

MAP for weights

Posterior Predictive for next output

What is so special about the Gaussian?

Given

Two independent Gaussian r.v.s

$$x \sim \mathcal{N}(\mu_x, \sigma_x^2)$$
$$y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

Can Show That

Their joint is Gaussian

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

Two linearly-dependent Gaussian r.v.s

$$x \sim \mathcal{N}(\mu_x, \beta_x^{-1})$$
$$y \sim \mathcal{N}(mx+b, \beta_y^{-1})$$

key: y only depends on x via mean which is linear func. of x

Their joint is Gaussian

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ m\mu_x + b \end{bmatrix}, \begin{bmatrix} \beta_x + \beta_y m^2 & -\beta_y m \\ -\beta_y m & \beta_y \end{bmatrix}^{-1} \right)$$

See PRML 2.103 and 2.113

A joint distribution over a partitioned vector

$$[-x-]^T = [-x_A-|-x_B-]^T$$

$$\begin{bmatrix} x_A \\ x_B \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \left[\begin{array}{c|c} \Sigma_{AA} & \Sigma_{AB} \\ \hline \Sigma_{AB}^T & \Sigma_{BB} \end{array}\right] \right)$$

Marginal is Gaussian PRML Eq. 2.98

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

Conditional is Gaussian PRML Eq. 2.96

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

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

Two linearly dependent Gaussians

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

Posterior is Gaussian

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

see formula in PRML Eq. 2.116

From Joint Gaussian Distribution to Marginal + Conditional

We have vector $X = [x_1, x_2, x_3, \ldots x_{D-1}, x_D]$ in D dimensions

Consider any exclusive partition of indices into A and B

e.g. if $D = 4$, then $A = \{1\}$, $B = \{2, 3, 4\}$

or

$A = \{2, 3\}$, $B = \{1, 4\}$

Define $X \sim \mathcal{N}(\mu, \Sigma)$ assuming we've reordered dims so we count first thru A then B

this means: $\begin{bmatrix} x_A \\ x_B \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} \right)$

$D \times 1$ vector

$D \times D$ cov matrix

Section 2.3.1 + 2.3.2

Bishop has lots of math deriving the following:

What is the conditional $p(x_A | x_B)$?

What is the marginal $p(x_A)$?

$p(x_B)$?

You should understand the math, but we'll skip math and show results

From Joint Gaussian to Marginal + Conditional

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

whole covariance matrix Σ is $D \times D$

has an inverse $\Lambda = \Sigma^{-1}$

$$= \begin{bmatrix} \Lambda_{AA} & \Lambda_{AB} \\ \Lambda_{BA} & \Lambda_{BB} \end{bmatrix}$$

Marginal

$$p(x_A) = \int p(x_A, x_B)\, dx_B$$

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

Conditional

$$p(x_A \mid x_B) = \frac{p(x_A, x_B)}{p(x_B)}$$

$$= N\left(x_A \mid \mu_A - \Lambda_{AA}^{-1} \Lambda_{AB} (x_B - \mu_B),\ \Lambda_{AA}^{-1} \right)$$

dim check: $A \times 1 - (A \times A)(A \times B)(B \times 1 - B \times 1),\ A \times A$

$= A \times 1$ ✓ ✓

The Gaussian-Gaussian model for regression

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

Prior

$$p(w) = \mathcal{N}(m_0, S_0)$$

$m_0 \in \mathbb{R}^M$

S_0 is $M \times M$ covariance matrix (symmetric, pos. def.)

Likelihood

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

↖ linear function of w!

iid each example

Turns out, posterior is given by (see Bishop 2.116)

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

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

$$S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi \quad \text{Eq. 3.50}$$

Check: dimensions,

what if no data?

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 Gaussian: $N(m_N, S_N)$

thus,

$$w_{MAP} = m_N$$

Why? Use property that mode of any Gaussian is its mean parameter

$$= \left(S_0^{-1} + \beta \Phi^T \Phi\right)^{-1} \left(S_0^{-1} m_0 + \beta \Phi^T t\right)$$

Consider a prior that favors zero mean + diag. covariance

$$m_0 = \vec{0} \quad \text{and} \quad S_0 = \alpha^{-1} I_M = \begin{bmatrix} \alpha^{-1} & & \\ & \alpha^{-1} & \\ & & \ddots \end{bmatrix}$$

$$S_0^{-1} = \alpha I_M$$

then our MAP simplifies to

$$w_{MAP} = \left(\alpha I_M + \beta \Phi^T \Phi\right)^{-1} \beta \Phi^T t$$

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

factor $\frac{1}{\beta}$ out of inverse

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

Looks familar! Equivalent to our sum-of-squares penalized linear regression w/ $\lambda = \frac{\alpha}{\beta}$

Predictive distribution for Bayes linear regression

See PRML Sec. 3.3.2

Assume: Linear Regression model with prior $m_0 = \vec{0}$, $S_0 = \alpha^{-1} I$, precision β known

Goal: Probability of next output t_* given its input features x_* and all training data $\{x_n, t_n\}_{n=1}^N$

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

also condition on $\alpha > 0$, $\beta > 0$

$p(t_* \mid x_*, w, \beta)$: likelihood — this is Gaussian, with mean linear in w

$p(w \mid \{x_n, t_n\}_{n=1}^N, \alpha, \beta)$: posterior given train data — this is Gaussian for w

Applying our Gaussian rules, looking for marginal $p(t_*)$ given joint Gaussian $p(w, t_*)$, we have

$$p(t_* \mid x_*, \{x_n, t_n\}_{n=1}^N) = \mathcal{N}\left(t_* \mid m_N^T \phi(x_*), \sigma_N^2(x_*)\right)$$

scalar! scalar mean scalar variance

where

$$\sigma_N^2(x_*) = \frac{1}{\beta} + \phi(x_*)^T S_N \phi(x_*)$$

PRML Eq 3.59

Key Message: Has simple form given posterior

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