Bayesian Linear Regression

data: observe pairs $\{x_i, y_i\}_{i=1}^{N}$

goals: given new x_*, predict its y_* (or sample plausible y_*)

across many possible x, what is shape of $y(x)$?

model:

$$y_i = y(x_i) = \underbrace{f(x_i)}_{\text{deterministic mean}} + \underbrace{\varepsilon_i}_{\text{"noise"}} \qquad \varepsilon_i \sim N(0, \sigma_n^2)$$

n for noise

Assume a linear model for mean function

$$f(x_i) = w * x_i \quad \text{in 1D}$$
$$= w^T x_i \quad \text{when } x_i \text{ is a D-dim vector}$$

Picture

several possible values for slope w are likely/plausible

Assume Gaussian prior on weights w

$$p(w) = \begin{cases} w \sim N(0, \sigma_p^2) & \text{in 1D} \\ \vec{w} \sim N(0, \Sigma_p) & \text{in D-dimensions} \end{cases}$$

Goal:

what is $p(w|x,y)$?

Bayes thm says:

$$p(w|x,y) = \frac{p(y|x,w)\, p(w)}{p(y|x)} = \frac{\text{lik} * \text{prior}}{\text{marg. lik.}}$$

1D posterior computation

$$p(w|y,x) \propto \overset{\text{const}}{\cancel{\frac{1}{p(y|x)}}}\; p(w) \prod_{i=1}^{N} p(y_i|x_i, w)$$

$$\propto \mathcal{N}(w|0, \sigma_p^2) \prod_{i=1}^{N} \mathcal{N}(y_i | w \cdot x_i, \sigma_n^2)$$

$$\propto \exp\left\{-\frac{1}{2}\frac{1}{\sigma_p^2} w^2\right\} \prod_{i=1}^{N} \exp\left\{-\frac{1}{2}\frac{1}{\sigma_n^2}(y_i - wx_i)^2\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_p^2} w^2 + \frac{1}{\sigma_n^2}\sum_{i=1}^{N}(y_i - wx_i)^2\right)\right\}$$

$$\propto \left(\frac{1}{\sigma_p^2} w^2 + \frac{1}{\sigma_n^2}\left[\overset{\text{const}}{\cancel{\sum_i y_i^2}} - 2\sum_i y_i w x_i + w^2 \sum_i x_i^2\right]\right)$$

$$\propto \exp\left\{-\frac{1}{2}\underbrace{\left(\frac{1}{\sigma_p^2} + \frac{1}{\sigma_n^2}\sum_i x_i^2\right)}_{a} w^2 - \underbrace{\left(\frac{1}{\sigma_n^2}\sum_i y_i x_i\right)}_{b} w\right\}$$

$$\propto \exp\left\{-\frac{1}{2} a w^2 - b w\right\}$$

"complete the square"
add $-\frac{1}{2}ab^2$ which is const wrt w

$$\propto \exp\left\{-\frac{1}{2} a \left(w - \frac{b}{a}\right)^2\right\}$$

looks like Gaussian pdf! $\propto \exp\left\{-\frac{1}{2}\frac{1}{\sigma^2}(w-\mu)^2\right\}$

$$\sigma^2 = \frac{1}{a} = \left(\frac{1}{\sigma_p^2} + \frac{1}{\sigma_n^2}\sum x_i^2\right)^{-1} \qquad \mu = \frac{b}{a} = \frac{\frac{1}{\sigma_n^2}\left(\sum_i y_i x_i\right)}{\frac{1}{\sigma_p^2} + \frac{1}{\sigma_n^2}\sum x_i^2}$$

Key Idea: Gaussian prior + Gaussian likelihood → Gauss posterior

Assuming model is good, we now can do closed form predictions

Compute exact samples for this posterior

General multivariate posterior with $D > 1$ dims

$$p(w) = \mathcal{N}_D(w \mid 0, \Sigma_p)$$

$$p(y_i \mid x_i, w) = \mathcal{N}_1(y_i \mid w^T x_i, \sigma_n^2) \qquad (y_i \text{ still scalar})$$

We apply same process

$$p(w \mid y, x) \propto \underbrace{\cancel{\frac{1}{p(y|x)}}}_{\text{const}} p(w) \prod_{i=1}^{N} p(y_i \mid x_i, w)$$

$$\propto \mathcal{N}(w \mid 0, \Sigma_p) \prod_i \mathcal{N}(y_i \mid w^T x_i, \sigma_n^2)$$

$$\propto \exp\left\{-\frac{1}{2} w^T \Sigma_p^{-1} w\right\} \exp\left\{-\frac{1}{2}\frac{1}{\sigma_n^2}(\vec{y} - X^T\vec{w})^T(\vec{y} - X^T\vec{w})\right\}$$

X: $D \times N$ matrix

$$\begin{bmatrix} | & & | & & | \\ x_1 & \cdots & x_i & \cdots & x_N \\ | & & | & & | \end{bmatrix}$$

$$\propto \exp\left\{-\frac{1}{2}(w - \mu)^T \Sigma^{-1} (w - \mu)\right\}$$

Posterior: $p(w \mid y, x) = \mathcal{N}(w \mid \mu, \Sigma)$

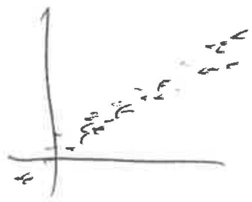
$$A = \Sigma^{-1} = \Sigma_p^{-1} + \frac{1}{\sigma_n^2} X X^T$$

$$\mu = \frac{1}{\sigma_n^2} A^{-1} X \vec{y}$$

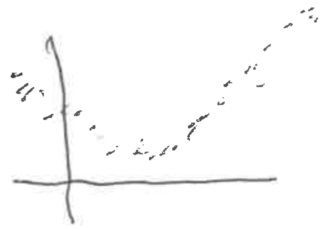
$\boxed{\text{Concept check}}$
Does it match 1D case?
Do dims work?
What is posterior if no data observed?

$\boxed{\text{Beyond Linear Features}}$ Kernel Trick

In previous ML course, should have seen more flexible features



linear OK



linear will FAIL

Possible feature maps:

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix} \quad \text{"quadratic"} \quad \text{or} \quad \begin{bmatrix} 1 \\ x \\ \cos(x) \\ \sin(x) \end{bmatrix} \quad \text{"periodic"}$$

New regression model

$$y(x_i) = \underbrace{w^T \phi(x_i)}_{\text{linear in FEATURE space}} + \underbrace{\varepsilon_i}_{\text{noise}} \qquad \phi(x_i) = \underbrace{\begin{bmatrix} \vdots \end{bmatrix}}_{F \text{ dim feature space}}$$

Now, can write posterior as

w also has length F

$$p(w \mid y, x) = \mathcal{N}\left(w \,\middle|\, \frac{1}{\sigma_n^2} A^{-1} \Phi \vec{y},\ A^{-1} \right)$$

where $A = \Sigma_p^{-1} + \frac{1}{\sigma_n^2} \Phi \Phi^T$

Check how big is A?

Can write posterior predictive as

$$p(f_* \mid x_*, x, y) = \mathcal{N}\left(f_* \,\middle|\, \frac{1}{\sigma_n^2} \phi(x_*)^T A^{-1} \Phi y,\ \phi(x_*)^T A^{-1} \phi(x_*) \right)$$

What values will function take at new test points?

Day 1 / 05

Consider a neat way to rewrite

A has size $F \times F$

$$\sigma^2_{\text{posterior}} = \phi(x_*)^T A^{-1} \phi(x_*)$$

$$= \phi(x_*)^T \left(\Sigma_p^{-1} + \frac{1}{\sigma_n^2} \Phi \Phi^T\right)^{-1} \phi(x_*)$$

$K = \Phi^T \Sigma_p \Phi$ is $N \times N$

no inverse here

$$= \phi(x_*)^T \Sigma_p \phi(x_*) - \phi(x_*)^T \Sigma_p \Phi \left(K + \sigma_n^2 I\right)^{-1} \Phi^T \Sigma_p \phi(x_*)$$

USING MATRIX INVERSION LEMMA

when $N < F$, cheaper to solve using "K" formula
this is the kernel trick, never need to fully represent feature vectors

define function

$$k(x, x') = \phi(x)^T \Sigma_p \phi(x')$$

we'll call this a kernel function
aka covariance function

output: scalar, > 0
larger values $\Rightarrow$ x and x' are closely correlated
zero $\Rightarrow$ x, x' not related

$$\sigma^2_{\text{posterior}} = k(x_*, x_*) - k(x_*, X)^T \left[k(X, X) + \sigma_n^2 I_N\right]^{-1} k(X, x_*)$$

$$\mu_{\text{posterior}} = \underset{1 \times N}{k(x_*, X)^T} \left[\underset{N \times N}{k(X, X)} + \underset{N \times N}{\sigma_n^2 I_N}\right]^{-1} y_{N \times 1}$$

Day 1/06

For Next Time: Gaussian Process

Read R & W Ch. 2, esp. 2.2 which will be in-class focus

submit critical comments to Canvas!

GPs let us take advantage of kernel trick to use very flexible feature embeddings $\phi(x)$ that might be infinite in dimension

GPs use the Gaussian and its useful properties

- conjugacy
- marginalization

if $p\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \middle| \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{matrix} \Sigma_{11} & \\ & \Sigma_{22} \end{matrix}\right)$

then $p(x_1) = \mathcal{N}(x_1 \mid \mu_1, \Sigma_{11})$

Seen: Weights view

prior is on vector w

$w \sim p(w)$

$y|x \sim p(y \mid w^T\phi(x), \sigma_n^2)$

New: Function view

prior is over possible function values

specify $m(x) = \mathbb{E}[f(x)]$

$k(x,x') = \mathrm{Cov}[f(x)]$

$$\begin{array}{c|c} f_1 & x_1 \\ f_2 & \vdots \\ \vdots & \vdots \\ f_N & x_N \end{array} \sim GP(\ )$$

$\sim \mathcal{N}(\underline{\text{mean}}, \underline{\text{covar}})$