Linear Regression

2/19/2013

w world state
x observed data

$$w_i \sim N(\phi_0 + \phi_1 x_i, \sigma^2)$$

redefine
$\phi = \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix}$ $x_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$

$\Theta = \{\phi_0, \phi_1, \sigma\}$

$$p(w_i | x_i, \theta) = N(x_i^T \phi, \sigma^2)$$

assume (w_i, x_i) Training examples are independent

Then define $1 \times N$ matrix $X = [x_1 \; x_2 \; \dots \; x_n]$

$N \times 1$ vector $w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{pmatrix}$

$I_N = N \times N$ Identity matrix

$$p(w | X, \Theta) = N(X^T \phi, \sigma^2 I_N)$$

$$= \frac{1}{(2\pi)^{N/2} |\det \sigma^2 I_N|^{1/2}} \exp\left\{-\frac{1}{2}(w - X^T\phi)^T (\sigma^2 I)^{-1} (w - X^T\phi)\right\}$$

$$= (2\pi)^{-N/2} \sigma^{-N} \exp\left\{-\frac{1}{2\sigma^2}(w - X^T\phi)^T (w - X^T\phi)\right\}$$

$$\log p = -\frac{N}{2}\log 2\pi - N \log \sigma - \frac{1}{2\sigma^2}\underbrace{(w - X^T\phi)^T (w - X^T\phi)}$$

$\underbrace{w^T w - 2\phi^T X w + \phi^T X X^T \phi}$ call it "M"

$$\frac{\partial \log p}{\partial \phi} = -2Xw + 2XX^T\phi = 0$$

$$XX^T\phi = Xw \rightarrow \phi = (XX^T)^{-1} X w$$

$$\frac{\partial \log p}{\partial \sigma} = -\frac{N}{\sigma} - \frac{1}{2} M(-2\sigma^{-3}) = 0$$

$$\frac{M}{\sigma^3} = \frac{N}{\sigma} \qquad \sigma^2 = \frac{M}{N} = \frac{(w - X^T\phi)^T (w - X^T\phi)}{N}$$

Linear Regression Cont'd

2/21/13

Recall w_i = world state (1×1), x_i = observed data (D×1)

$$P(w_i \mid x_i, \phi, \sigma^2) = N(x_i^T \phi, \sigma^2)$$

want to learn ϕ, σ^2 from training data pairs (x_i, w_i)

$X = [x_1, x_2, \ldots, x_N]$ (D×N matrix) $\quad w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}$ (N×1 vector) $\quad I_N$ = N×N Identity matrix

MLE estimation

$$\Rightarrow \quad \hat{\phi} = (XX^T)^{-1} X w \qquad \hat{\sigma}^2 = \frac{(w - X^T\hat{\phi})^T (w - X^T\hat{\phi})}{N}$$

Predictive distribution: given a new measurement x^*, compute the posterior distribution over values of w

$$P(w \mid x^*, \hat{\phi}, \hat{\sigma}^2) = N(x^{*T}\hat{\phi}, \hat{\sigma}^2)$$
$$\propto \exp\left\{ -\frac{1}{2\hat{\sigma}^2} (w - x^{*T}\hat{\phi})^2 \right\}$$

We would like to take a Bayesian approach in most cases, to "regularize" the fitting (via a prior on ϕ) and thus avoid overfitting. This is especially important as data dimension D gets larger

Introduce a prior on ϕ

$$p(\phi) = \mathcal{N}(\phi \mid 0, s^2 I_d)$$

zero mean
variance s^2 (in general a different variance than σ^2)

note: s^2 should be relatively large, if we are initially uncertain about values of coefficients of ϕ.

note: making this prior zero-mean influences estimates of ϕ to have smaller ~~coeffs~~ magnitude. This is what has the regularizing (smoothing) effect.

By the way, the Bayesian approach to regression we are deriving, using this prior, is equivalent to "Ridge regression".

using Bayes Rule to compute a posterior over ϕ

Dimensions
$w = N \times 1$
$x_i = D \times 1$
$X = D \times N$
$\phi = D \times 1$
$I_N = N \times N$
$I_d = D \times D$

$$p(\phi \mid x, w) = p(w \mid x, \phi)\, p(\phi) \,/\, p(w \mid x)$$

$$\propto \mathcal{N}(w \mid X^T\phi, \sigma^2 I_N)\, \mathcal{N}(\phi \mid 0, s^2 I_d)$$

$$\propto \exp\left\{ -\frac{1}{2\sigma^2}(w - X^T\phi)^T(w - X^T\phi) - \frac{1}{2s^2}\phi^T I_d \phi \right\}$$

$$= \exp\left\{ -\frac{1}{2}\left[\frac{1}{s^2}\phi^T I_d \phi + \frac{w^T w}{\sigma^2} - \frac{2\phi^T X w}{\sigma^2} + \frac{1}{\sigma^2}\phi^T X X^T \phi \right] \right\}$$

$$= \exp\left\{ -\frac{1}{2}\left[\phi^T \overbrace{\left(\frac{1}{s^2} I_d + \frac{1}{\sigma^2} X X^T\right)}^{A} \phi - 2\phi^T \overbrace{\left(\frac{1}{\sigma^2} X w\right)}^{b} + \overbrace{\frac{1}{\sigma^2} w^T w}^{c} \right] \right\}$$

Fun Fact

any distribution proportional to $\exp\{-\frac{1}{2}[ax^2 - 2bx + c]\}$ for a scalar variable is a Gaussian with mean $\frac{b}{a}$ and variance $\frac{1}{a}$!

Proof: "completing the square" - notes will be posted on our web site.

matrix version of + is a vector variable

$$\exp\{-\tfrac{1}{2}[x^T A x - 2x^T b + c]\}$$

is a Gaussian with mean $A^{-1}b$ and variance A^{-1}

[note A is called a "precession matrix".]

So $p(\phi | x, w)$

$$\propto \exp\{-\tfrac{1}{2}\phi^T A \phi - 2\phi^T b + c\}$$

$$A = \frac{1}{s^2} I_d + \frac{1}{\sigma^2} x x^T$$
$$b = \frac{1}{\sigma^2} x w$$
$$c = \frac{1}{\sigma^2} w^T w$$

is Gaussian with

mean $\frac{1}{\sigma^2} A^{-1} x w$ ($D \times 1$) and variance A^{-1} ($D \times D$)

where A ($D \times D$) $= \left(\frac{1}{\sigma^2} x x^T + \frac{1}{s^2} I_d\right)$

We could now compute a MAP estimate for ϕ to maximize this. It will of course be the mean of this Gaussian posterior

$$\hat{\phi}_{map} = \frac{1}{\sigma^2} A^{-1} x w$$

Alternatively, we could do full Bayesian inference for a new ~~value~~ measurement x^* to compute the distribution of values over predicted world state w

Recall this is essentially computing the expected value of $p(w|x^*, \phi)$ with respect to the posterior distribution $p(\phi|x, w)$

$$p(w|x^*, x, w) = \int_\phi p(w|x^*, \phi)\, p(\phi|x, w)\, d\phi$$

$$= \int \mathcal{N}(w|x^{*T}\phi, \sigma^2)\, \mathcal{N}(\phi|\tfrac{1}{\sigma^2}A^{-1}Xw, A^{-1})\, d\phi$$

IT TURNS OUT THAT THIS IS ALSO A GAUSSIAN,

$$= \mathcal{N}\Big(w|\underbrace{\tfrac{1}{\sigma^2}\cdot x^{*T}A^{-1}Xw}_{\substack{1\times 1\\ \text{mean}}},\ \underbrace{x^{*T}A^{-1}x^* + a}_{\substack{1\times 1\\ \text{variance}}}\Big)$$