Bayesian Regression

Mon Apr 12

Linear Discriminant Analysis

<u>Regression</u>

<u>Discriminative model</u> $p(y|x)$

vs.

<u>Generative Model</u>

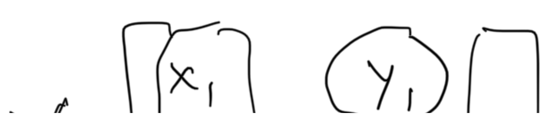
$p(y)$ $p(x|y)$

---

w — ?

$$y|x,w = \mathcal{N}(w^T x, \sigma^2)$$

scalar



$$D = \begin{array}{c} \\ \end{array}\begin{array}{cc} X & Y \\ \left[\begin{matrix} x_2 \\ x_3 \\ \vdots \\ x_i \\ \vdots \\ x_N \end{matrix}\right. & \left.\begin{matrix} y_2 \\ y_3 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{matrix}\right] \end{array}$$

$$L(D|w) = \prod_{i=1}^{N} \underline{p(y_i | x_i; w)}$$

$$LL(D|w) = \sum_{i=1}^{N} \log \underline{p}(y_i | x_i, w)$$

$$= \sum_{i=1}^{N} \log\left[\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - w^T x_i)^2\right]\right]$$

$$= \sum_{i=1}^{N}\left[-\log(\sqrt{2\pi}\,\sigma) - \frac{1}{2\sigma^2}(y_i - w^T x_i)^2\right]$$

$$= \underbrace{-\frac{N}{2}\log 2\pi - N\log\sigma}\; - \frac{1}{2\sigma^2}\underline{\sum_{i=1}^{N}(y_i - w^T x_i}$$

Find (w, σ^2) that maximizes $LL(D|w)$

$$\frac{\partial LL(D|w)}{} = 0 \quad \Big| \quad \frac{\partial LL(D|w)}{} = 0$$

$\partial w \qquad\qquad | \qquad \partial \sigma$

$$\hat{w}_{MLE} = (X^T X)^{-1} X^T y$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{N} (y - X\hat{w}_{MLE})^T (y - X\hat{w}_{MLE})$$

Putting a prior on $\mathbf{w}$

$\mathbf{w}$ is a $(D+1)$ length vector

$$p(\mathbf{w}) \sim \mathcal{N}(w \mid \mu_0, \Sigma_0)$$

$$p(w \mid D) = \frac{p(D \mid w)\, p(w)}{\int_{w'} p(D \mid w')\, p(w')\, dw'}$$

Posterior ~~wil~~ of $\mathbf{w}$ will also be a Gaussian.

$$p(w) \sim \mathcal{N}(w \mid 0, \tau^2 I)$$

scalar

Special but often-used prior on w

Posterior:

$$\boxed{\begin{aligned} \bar{W} &= \left(X^T X + \frac{\sigma^2}{\tau^2} I\right)^{-1} X^T y \\ \bar{\Sigma} &= \sigma^2\left(X^T X + \frac{\sigma^2}{\tau^2} I\right)^{-1} \end{aligned}}$$

Think ridge regression

$$y \sim \mathcal{N}(w^T x, \sigma^2)$$

Generalized Linear Model

$$pdf(y|x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\frac{(y_i - w^T x)^2}{\sigma^2}\right]$$

$$Laplace(y|x) = \frac{1}{-} \exp \|y - w^T x_i\|$$

Estimating w using Laplace distribution is more robust!

challenging.

$\underline{y \mid x} \sim \text{Bernoulli}(\theta)$

$0 \leq \theta \leq 1$

Binary classification.

$$\theta = \sigma(w^T x)$$

$$= \frac{1}{1 + \exp(-w^T x)}$$

If $y_i = 1$ $\quad p(y_i) = \theta_i$

$$= \frac{1}{1 + \exp(-w^T x_i)}$$

If $y_i = 0$ $\quad p(y_i) = 1 - \theta_i$

$$= \frac{1}{1 + \exp(w^T x_i)}$$

In general

$$p(y_i) = \frac{1}{1 + \exp(-y_i w^T x_i)}$$

$$L(w) = \prod^{N} \left[\frac{1}{\;\cdots\;} \right]$$

$$\prod_{i=1} [1+\exp(-y_i w^T x_i)]$$

$$\boxed{LL(w) = -\sum_{i=1}^{N} \log\left(1+\exp(-y_i w^T x_i)\right)}$$

No closed-form solution

Gradient Descent

Regularizing LR

$$w \sim \mathcal{N}(0, \tau^2 I)$$

$$p(w|X,y) \propto p(w)\, p(X,y|w)$$

Multinoulli

$$X \in \{1, 2, ---, C\}$$

$$\theta = [0.1, 0.2, 0.05, \ldots\ldots 0.2]$$

$$\theta_j = \frac{\exp(w_j^T X)}{\quad}$$

$$\sum_{k=1}^{C} \exp(w_k^T x)$$