Tutorial for A6

Thursday, October 30, 2014 3:01 PM

There will be a tutorial next week!

Conjugate function

$$f^*(y) = \sup_x \{ y^T x - f(x) \}$$

Derivation for these? Check section 3.3

$$f(x) \xrightarrow{\text{conj}} f^*(y)$$

$$h(x) = f(Ax + b)$$

$$h(x) \xrightarrow{\text{conj}} h^*(y) = f^*(A^T y) - b^T A^{-T} y$$

A^{-T}: inverse and transpose

Example:

getting conjugate of function:

$$f(x) = x \log x \qquad \text{Domain: } x > 0$$

$$f^*(y) = \sup_x \{ y^T x - x \log x \}$$

d/dx, set to 0

$$y - \log x - 1 = 0 \qquad \text{Domain of } f^* = \mathbb{R}$$

$$\log x = y - 1$$

$$x = e^{y-1} \Rightarrow \text{value for max}$$

$$f^*(y) = e^{y-1}(y - y + 1) = e^{(y-1)}$$

$$P(x) = f(Ax) + g(x)$$

$$D(y) = -\left[f^*(y) + g^*(A^T y) \right]$$

(important)

- why dual? sometimes its easier to solve and can give same end solution (under conditions)

1.1.1

$$P(x) = \tfrac{1}{2} \|Ax - b\|^2 + \tfrac{\lambda}{2} \|x\|^2$$

$D(y) = ?$

$\|Ax - b\|^2$: f $\qquad$ $\frac{\lambda}{2}\|x\|^2$: g

dual of this is the same! (from sec 3.3)

We need f in the form $F(Ax)$

start without A:

$$F(x) = \tfrac{1}{2} \|x - b\|^2$$

$$F^*(y) = \tfrac{1}{2} \|y\|^2 + b^T y$$

$$g^*(y) = \tfrac{1}{2\lambda} \|y\|^2$$

$f(x) = F(Ax)$

$f(x) = \frac{1}{2}\|x\|^2$

$f^*(y) = \frac{1}{2}\|y\|^2$

$-b^T I^{-T} y$ (I = identity)

$$h(x) \xrightarrow{\text{conj}} h^*(y) = f^*(A^T y) - b^T A^{-T} y$$

$$D(y) = -\left[\tfrac{1}{2} \|y^2\| + b^T y + \frac{1}{2\lambda} \|A^T y\|^2 \right] \quad \Leftarrow \text{solution}$$

1.1.2

$$f(x) = \|x\| \xrightarrow{*\text{ conj}} f^*(y) = \begin{cases} 0 & \text{if } \|y\|_* \leq 1 \\ \infty & \text{else} \end{cases}$$

dual norm

} defines the domain of f^*

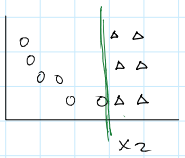
1.2 Coordinate ascent:

Need to project solution on interval $0 \to 1$

2.1 data is currently replicated all the same, just change to bootstrapping.

(sampling with replacement, m different models to train, average the result.)

2.2

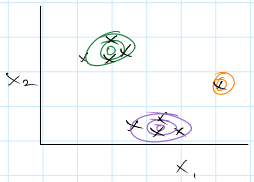


decision stump threshold

find dimension and threshold that minimizes error

(just pick the best one)

3.1 See section: 11.4.2



fit k gaussians

(gaussians can overlap)

Code:
change to work with k gaussians, and probability coresponds to each gaussian

Initialize $\theta^0 \left[\pi_k, \mu_k, \Sigma_k \right]$ ← weight / prior of each gaussian

E-step $p(y_i = k \mid x_i, \theta^{t-1}) = r_{ik}$ belongs to gaussian k with some probability

$$= \frac{\pi_k \, p(x_i \mid \theta_k^{t-1})}{\sum_{k'} \pi_{k'} \, p(x_i \mid \theta_{k'}^{t-1})}$$

M step

$$\pi_k = \frac{1}{N} \sum_i r_{ik}$$

$$\mu_k = \frac{\sum_i r_{ik} \, x_i}{r_k}$$

$$\Sigma_k = \frac{\sum_i r_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{r_k}$$

3.2

log likelihood — features, labels

$$LL(\theta) = \log P(X, Y \mid \theta)$$

dataset: (X, Y) → (X_L, Y_L, X_U) (missing Y_U)

$$= \log \prod_{i=1}^{N} p(x_i, y_i \mid \theta)$$

joint distribution ← condition on y — features indep. given labels

$$= \sum_{i=1}^{N} \log p(x_i \mid y_i, \theta) + \log p(y_i \mid \theta)$$

$\prod_{j=1}^{d} p(x_{ij} \mid y_i, \theta)$ naive bayes assumption

$$Q(\theta \mid \theta^{t-1}) = E\left[LL(\theta)\right]$$

$p(Y_U \mid X, \theta^{t-1})$

Example initialization: $\theta = [$ vector $]$

$p(y_i = 1) = 1/2$

$p(y_i = 0) = 1/2$

$$Q(\theta \mid \theta^{t-1}) = \sum_{i=1}^{N} \log p(x_i, y_i) + \sum_{i=1}^{N_T} \sum_{y=\{0,1\}} p(y_i \mid x_i, \theta^{t-1}) \log p(x_i, y_i)$$

N_T ← unlabeled points

labeled — unlabeled data

take derivative, equate to zero

$$\pi_1 = p(y = 1) = \frac{\sum_{i=1}^{N} I(y_i = 1)}{N} \quad \text{(for complete data)}$$

roughly speaking: $\Rightarrow \frac{\sum_{i=1}^{N} I(y_i = 1)\, w_i}{N}$ ← weight of unlabeled data