2 Multi-Class Classification

Problem 2: Consider a generative classification model for C classes defined by prior class probabilities $p(y = c) = \pi_c$ and general class-conditional densities $p(\boldsymbol{x}|y = c, \boldsymbol{\theta}_c)$ where $\boldsymbol{x}$ is the input feature vector and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_c\}_{c=1}^{C}$ are further model parameters. Suppose we are given a training set $\mathcal{D} = \{(\boldsymbol{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$ where $y^{(n)}$ is a binary target vector of length C that uses the 1-of-C (one-hot) encoding scheme, so that it has components $y_c^{(n)} = \delta_{ck}$ if pattern n is from class $y = k$. Assuming that the data points are iid, show that the maximum-likelihood solution for the prior probabilities is given by

$$\pi_c = \frac{N_c}{N}$$

where N_c is the number of data points assigned to class $y = c$.

K

To show: $\underset{\pi_c}{\operatorname{argmax}}\; p(D \mid \pi, \theta) = \frac{N_c}{N}$

$$p(D \mid \pi, \theta) = \prod_{n=1}^{N} \prod_{c=1}^{K} p(x^{(n)} \mid \theta_c)\, p(y_c^{(n)} \mid \pi_c)$$

$$L(\pi, \theta) = \sum_{n=1}^{N} \sum_{c=1}^{K} \underbrace{\log p(x^{(n)} \mid \theta_c)}_{\text{const in } \pi} + \log p(y_c^{(n)} \mid \pi_c)$$

$$p(y = c \mid \theta) = \theta_c$$

$$\prod_{c=1}^{K} \theta_c^{y_c^{(n)}}$$

$$= \sum_{n=1}^{N} \sum_{c=1}^{K-1} \log p(y_c^{(n)} \mid \pi_c) + \log p(y_K^{(n)} \mid \pi_K)$$

$$\pi_K = 1 - \sum_{c=1}^{K-1} \pi_c$$

$$= \sum_{n=1}^{N} \sum_{c=1}^{K-1} y_c^{(n)} \log \pi_c + y_K^{(n)} \log \left[1 - \sum_{i=1}^{K-1} \pi_i\right]$$

$$\frac{\partial}{\partial \pi_c} L(\pi, \theta) \overset{!}{=} 0$$

$$\Leftrightarrow \sum_{n=1}^{N} \frac{y_c^{(n)}}{\pi_c} - \sum_{n=1}^{N} \frac{y_K^{(n)}}{1 - \sum_{i=1}^{K-1} \pi_i}$$

$$\frac{N_c}{\pi_c} = N_K \frac{1}{1 - \sum_{i=1}^{K-1} \pi_i}$$

plug in: $\pi_c = \frac{N_c}{N}$

$$N = N_K \cdot \frac{1}{1 - \sum_{i=1}^{K-1} \frac{N_i}{N}}$$

$$N - \sum^{K-1} N_i = N_K$$

$i=1$

$$N = N_k + \sum_{i=1}^{k-1} N_i$$

$$N = \underbrace{\sum_{i=1}^{k} N_i}_{N}$$

Linear Classification (2)

Tuesday, 13 November 2018 10:45

Problem 3: Using the same classification model as in the previous question, now suppose that the class-conditional densities are given by Gaussian distributions with a shared covariance matrix, so that

$$p(x|y=c,\boldsymbol{\theta}_c) = p(x|\boldsymbol{\theta}_c) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}).$$

Show that the maximum likelihood solution for the mean of the Gaussian distribution for class C_c is given by

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{\{n|\boldsymbol{x}^{(n)} \in C_c\}} \boldsymbol{x}^{(n)}$$

which represents the mean of those feature vectors assigned to class C_c.

N_c: Number of samples in class c

N: Number of samples

$x_i \in \mathbb{R}^d$

$$\mathcal{N}(x \mid \mu_c, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left[-\frac{1}{2}(x-\mu_c)^T \Sigma^{-1} (x-\mu_c)\right]$$

$$\log \mathcal{N}(x \mid \mu_c, \Sigma) = -\frac{1}{2}\left[\log|\Sigma| + (x-\mu_c)^T \Sigma^{-1} (x-\mu_c)\right] + \text{const in } \mu_c, \Sigma$$

$$p(\mathcal{D} \mid \mu, \Sigma) = \prod_{n=1}^{N} \prod_{c=1}^{C} p(x^{(n)} \mid \mu_c, \Sigma) \cdot p(y_c^{(n)})$$

-> Take the log

$$L(\mu, \Sigma) = \sum_{n=1}^{N} \sum_{c=1}^{C} \log p(x^{(n)} \mid \mu_c, \Sigma) + \underbrace{\log p(y_c^{(n)})}_{\text{const in } \mu, \Sigma}$$

$$= -\frac{1}{2} \sum_{n=1}^{N} \sum_{c=1}^{C} y_c^{(n)} \left[\overbrace{\log|\Sigma|}^{\text{const in } \mu_c} + (x^{(n)}-\mu_c)^T \Sigma^{-1} (x^{(n)}-\mu_c)\right]$$

$$\frac{\partial}{\partial \mu_c} L(\mu, \Sigma) \overset{!}{=} 0 \qquad \frac{\partial}{\partial x} x^T a = a = \frac{\partial}{\partial x} a^T x$$

* see detailed derivation below

$$0 = \sum_{n=1}^{N} y_c^{(n)} \Sigma^{-1} (x^{(n)} - \mu_c)$$

$$\mu_c = \frac{1}{N_c} \sum_{n=1}^{N} y_c^{(n)} x^{(n)} \checkmark$$

$$* \quad \frac{\partial}{\partial \mu_c} (x-\mu_c)^T \Sigma^{-1} (x-\mu_c) = \frac{\partial}{\partial \mu_c} \left[\overset{\text{const in } \mu_c}{x^T \Sigma^{-1} x} - x^T \Sigma^{-1} \mu_c - \mu_c^T \Sigma^{-1} x + \mu_c^T \Sigma^{-1} \mu_c\right]$$

$$\frac{\partial}{\partial \mu_c} x^T \Sigma^{-1} \mu_c = \frac{\partial}{\partial \mu_c} \mu_c^T \Sigma^{-1} x = \Sigma^{-1} x \qquad \frac{\partial}{\partial \mu_c} \mu_c^T \Sigma^{-1} \mu_c = \Sigma^{-1} \mu_c + \mu_c^T \cdot \Sigma^{-1} = 2\Sigma^{-1} \mu_c$$

$$= 2\Sigma^{-1} \mu_c - 2\Sigma^{-1} x$$

$$= 2\Sigma^{-1} (\mu_c - x)$$

Similarly, show that the maximum likelihood solution for the shared covariance matrix is given by

$$\Sigma = \sum_{c=1}^{C} \frac{N_c}{N} \mathbf{S}_c$$

where

$$\mathbf{S}_c = \frac{1}{N_c} \sum_{\{n | \boldsymbol{x}^{(n)} \in C_c\}} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_c)^T .$$

Thus Σ is given by a weighted average of the covariances of the data associated with each class, in which the weighting coefficients N_c/N are the prior probabilities of the classes.

To show: $\arg\max_{\Sigma} L(\mu, \Sigma) = \sum_{c=1}^{C} \frac{N_c}{N} S_c$

$$S_c = \frac{1}{N_c} \sum_{n=1}^{N} y_c^{(n)} \underbrace{\underset{d\times 1}{(x^{(n)} - \mu_c)} \underset{1\times d}{(x^{(n)} - \mu_c)^T}}_{d\times d}$$

$$L(\mu, \Sigma) = -\frac{1}{2} \sum_{n=1}^{N} \sum_{c=1}^{C} y_c \left[\log|\Sigma| + \underbrace{\underbrace{\underset{1\times d}{\overset{A}{(x^{(n)} - \mu_c)^T}} \underset{d\times d}{\overset{B}{\Sigma^{-1}}}}_{1\times d} \underset{d\times 1}{\overset{C}{(x^{(n)} - \mu_c)}}}_{1\times 1 \text{ (scalar)}} \right]$$

$$-\log|A| = \log|A^{-1}|$$

$$Tr(ABC) = Tr(BCA)$$

$Tr(a) = a$ for $a \in \mathbb{R}$

$$L(\mu, \Sigma) = -\frac{1}{2} \sum_{n=1}^{N} \sum_{c=1}^{C} y_c^{(n)} \left[\log|\Sigma^{-1}| - Tr\left[\Sigma^{-1} (x^{(n)} - \mu_c)(x^{(n)} - \mu_c)^T\right] \right]$$

$$\frac{\partial}{\partial A} \log|A| = (A^{-1})^T \qquad \frac{\partial}{\partial A} Tr(AB) = B^T$$

$$\frac{\partial}{\partial \Sigma^{-1}} L(\mu, \Sigma) = \frac{1}{2} \sum_{n=1}^{N} \sum_{c=1}^{C} y_c^{(n)} \left[\Sigma - (x^{(n)} - \mu_c)(x^{(n)} - \mu_c)^T \right] \overset{!}{=} 0$$

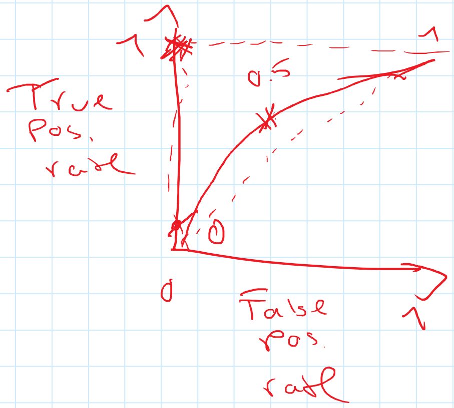
$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_c^{(n)} (x^{(n)} - \mu_c)(x^{(n)} - \mu_c)^T$$

$$= \sum_{c=1}^{C} \frac{N_c}{N} S_c \checkmark$$

$$S_c = \frac{1}{N_c} \sum_{n=1}^{N} y_c^{(n)} (x^{(n)} - \mu_c)(x^{(n)} - \mu_c)^T \checkmark$$

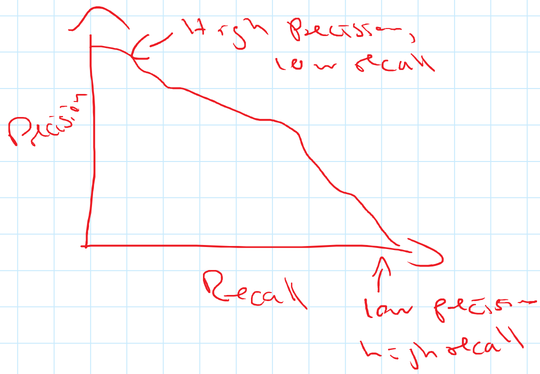
ROC curves

Receiver operator characteristic



Area under curve

Precision recall curve (PR curve) aka avg. precision



$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$