Machine Learning, Tutorial 4
University of Bern

Mehdi Noroozi (noroozi@inf.unibe.ch)

17/10/2018

Multivariate Gaussian

1. Prove the following statements regarding the covariance:

   (a) Show that if X and Y are independent then $\text{Cov}[X, Y] = 0$.
   Give an example that shows that the opposite is not true.
   Solution

   $$\begin{aligned}\text{Cov}[X, Y] &= E[(X - E[X])(Y - E[Y])] = E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] = E[XY] - E[X]E[Y]\end{aligned}$$

   If X, Y are independent, $E[XY] = E[X]E[Y]$ therefore the covariance is equal to 0. To show that the opposite is not correct, consider random variales $X \sim \mathcal{N}(0, 1), Y = X^2$ we know that $E(X) = E(X^3) = 0$. Threfore, $Cov(X, Y) = E[XY] - E[X]E[Y] = E[X^3] - E[X]E[X^2] = 0$. Howiver, it is clear that X, Y are not independent.

   (b) Show that the covariance matrix is always symmetric and positive semidefinite.
   Solution
   The $(i, j)^{th}$ element of the covariance matrix Σ is given by

   $$\Sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)] = E[(X_j - \mu_j)(X_i - \mu_i)] = \Sigma_{ji}$$

   so that the covariance matrix is symmetric.

   For an arbitrary vector u,

   $$\begin{aligned}u^\top \Sigma u = u^\top E[(X - \mu)(X - \mu)^\top]u &= E[u^\top (X - \mu)(X - \mu)^\top u] \\ = E[((X - \mu)^\top u)^\top (X - \mu)^\top u] &= E[((X - \mu)^\top u)^2] \geq 0\end{aligned}$$

   so that the covariance matrix is positive semidefinite.

2. For a function $f : R^2 \to R$, an isocontour is a set of the form

   $$\{x \in R^2 : f(x) = c\}$$

   for some $c \in R$.
   Derive an analytical form for the isocontours of a multivariate Gaussian.
   Solution The probability density function of the multivariate Gaussian distribution is

   $$p(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)). \tag{1}$$

   The isocontour:

   $$c = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)) \tag{2}$$

$$c(2\pi)^{n/2}|\Sigma|^{1/2} = exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)). \quad (3)$$

$$log(c) + \frac{n}{2}log(2\pi) + \frac{1}{2}log(|\Sigma|) = -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) \quad (4)$$

On the left side we have a constant. On the right site we have a quadratic function of x. Because Σ^{-1} is positive definite, the isocontour is an n dimensional ellipsoid.

3. Consider the classifier based on Gaussian Discriminant Analysis, where the distribution of the samples are modelled by

$$p(y) = \phi^y(1-\phi)^{(1-y)}, \quad (5)$$

$$p(x|y=0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)), \quad (6)$$

$$p(x|y=1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)). \quad (7)$$

Compute $p(y=1|x)$. How does this relate to Linear Regression?

Solution From the Bayes rule we get $p(y=1|x) = p(x|y=1)p(y=1)/p(x)$, similarly $p(y=0|x) = p(x|y=0)p(y=0)/p(x)$. The log-likelihood ratio becomes

$$\begin{aligned} L &= \log(p(y=1|x)/p(y=0|x)) = \\ &= \log(p(x|y=1)p(y=1)) - \log(p(x|y=0)p(y=0)) = \\ &= -\tfrac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \log(\phi) \\ &\quad +\tfrac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) - \log(1-\phi) = \\ &= W^T x + B, \end{aligned} \quad (8)$$

where the parameters $W = \Sigma^{-1}(\mu_1 - \mu_0)$ and $B = -\mu_1^T\Sigma^{-1}\mu_1 + \mu_0^T\Sigma^{-1}\mu_0 + \log(\phi) - \log(1-\phi)$.

The probability therefore becomes

$$p(y=1|x) = \frac{1}{1 + p(y=0|x)/p(y=1|x)} = \frac{1}{1+\exp(-(W^Tx+B))}. \quad (9)$$

This expression is the same as the linear regression, where $\theta = [W, B]$ can be thought as linear regression parameters.

Naive Bayes

1. Consider a text classification problem using the multinomial naive Bayes classifier. Given the following data we want to classify texts into two classes j(Japanese) and c(Chinese) based on the observed words.

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Give detailed answers to the following questions.

(a) Calculate the probabilities of the two classes $P(c)$ and $P(j)$.

(b) Calculate conditional probabilities $P(word|class)$ by using Laplacian smoothing.

(c) Write the inequality (with explicit number) used to classify the fifth document.

Solution.

$$
\begin{aligned}
P(c) &= \frac{3}{4} \\
P(j) &= \frac{1}{3}, \\
P(Chinese|c) &= (5+1)/(8+6) = \frac{3}{7} \\
P(Tokyo|c) &= (0+1)/(8+6) = \frac{1}{7} \\
P(Japan|c) &= (0+1)/(8+6) = \frac{1}{7} \\
P(Chinese|j) &= (1+1)/(3+6) = \frac{2}{9} \\
P(Tokyo|j) &= (1+1)/(3+6) = \frac{2}{9} \\
P(Japan|j) &= (1+1)/(3+6) = \frac{2}{9}, \\
P(c|d5) &\propto \frac{3}{4} * (\frac{3}{7})^3 * \frac{1}{14} * \frac{1}{14} \approx 0.0003 \\
P(j|d5) &\propto \frac{1}{4} * (\frac{2}{9})^3 * \frac{2}{9} * \frac{2}{9} \approx 0.0001
\end{aligned}
$$