# 582631 Introduction to Machine Learning, Fall 2016
# Exercise set III
# Model solutions

## 2.

(a) First, because $\mathbf{\Sigma}$ is a diagonal matrix, the square root of its determinant is

$$|\mathbf{\Sigma}|^{1/2} = \left(\prod_{i=1}^{p} \sigma_i^2\right)^{1/2} = \prod_{i=1}^{p} \sigma_i,$$

which is the term appearing in the constant in front of the exponent term. The inverse of a diagonal matrix is given by

$$\mathbf{\Sigma}^{-1} = \operatorname{diag}\left(\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_p^2}\right). \tag{1}$$

Thus, the result of the matrix product in the exponent term simplifies and gives

$$\begin{aligned}
(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= \begin{bmatrix} x_1-\mu_1 & \ldots & x_p-\mu_p \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \ldots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \frac{1}{\sigma_p^2} \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ \vdots \\ x_p-\mu_p \end{bmatrix} \\
&= \begin{bmatrix} \frac{x_1-\mu_1}{\sigma_1^2} & \ldots & \frac{x_p-\mu_p}{\sigma_p^2} \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ \vdots \\ x_p-\mu_p \end{bmatrix} \\
&= \sum_{i=1}^{p} \frac{(x_i-\mu_i)^2}{\sigma_i^2}.
\end{aligned} \tag{2}$$

Plugging (1) and (2) into the right-hand side of the given formula gives the left-hand side. Q.E.D.

(b) By plugging the density of the multivariate normal distribution into the equation we can solve $b$:

$$\begin{aligned}
\frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{b}{2}} &= a \\
e^{-\frac{b}{2}} &= (2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2} a \\
-\frac{b}{2} &= \log a + \frac{p}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{\Sigma}| \\
b &= -2\log a - p\log(2\pi) - \log|\mathbf{\Sigma}|.
\end{aligned}$$

The determinant of the covariance matrix is

$$|\mathbf{\Sigma}| = \sigma_1^2\sigma_2^2\sigma_3^2 = 6,$$

and by plugging this and $a = 1/100$, $p = 3$ into the formula above we get $b \approx 1.90495$.

**On the interpretation of $b$ (optional):** What is a geometric interpretation of $b$? Consider first a special case where the covariance matrix of the $p$-dimensional multinormal distribution is diagonal:

$$\boldsymbol{\Sigma}^{-1} = \operatorname{diag}\left(\frac{1}{\sigma_1^2}\cdots\frac{1}{\sigma_p^2}\right).$$

In the first part of the exercise we computed this matrix sum:

$$\begin{aligned} b &= (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\ b &= \sum_{i=1}^{p}\frac{(x_i-\mu_i)^2}{\sigma_i^2} \\ 1 &= \sum_{i=1}^{p}\left(\frac{x_i-\mu_i}{\sqrt{b}\,\sigma_i}\right)^2. \end{aligned}$$

This is an equation of the ellipsoid with a centerpoint

$$\boldsymbol{\mu} = (\mu_1,\ldots,\mu_p)$$

and semi-axes with lengths

$$\sqrt{b}\,\sigma_1,\ldots,\sqrt{b}\,\sigma_p.$$

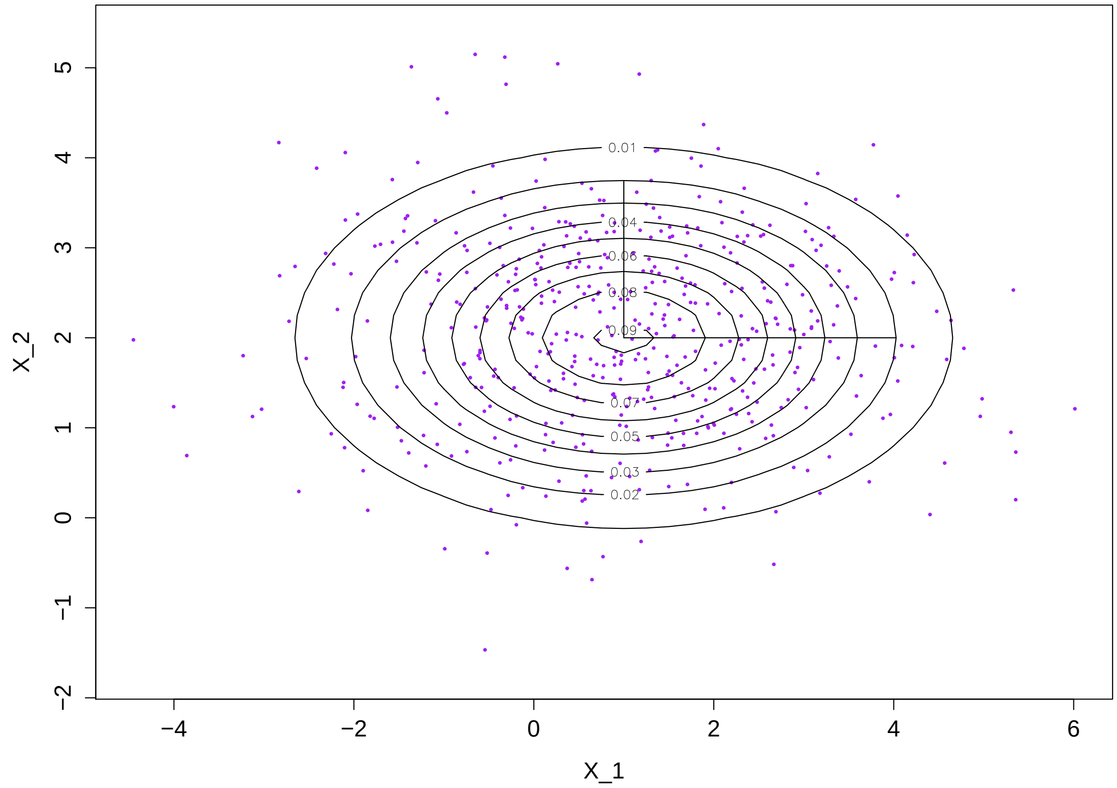

Figure 1: Diagonal covariance matrix with $\sigma_1^2 = 3$, $\sigma_2^2 = 1$.

For the two-dimensional case this is illustrated on Figure 1, which included data generated from the 2-dimensional multinormal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = (1, 1), \quad \boldsymbol{\Sigma} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix},$$

and contours of the density function of this distribution.

By computing $b$ corresponding to the contour of the density function where its value is $a = 0.02$ we get from the formula computed above $b \approx 3.05$; hence, the length of the semi-axes of this ellipse are

$$\sqrt{b \cdot 3} \approx 3.03, \quad \sqrt{b \cdot 1} \approx 1.75.$$

Line segments of these lengths are drawn along the coordinate axes starting from the centre of the ellipse $(1, 1)$. It can be seen that these are indeed the semi-axes of the contour of the density function with value $a = 0.02$.

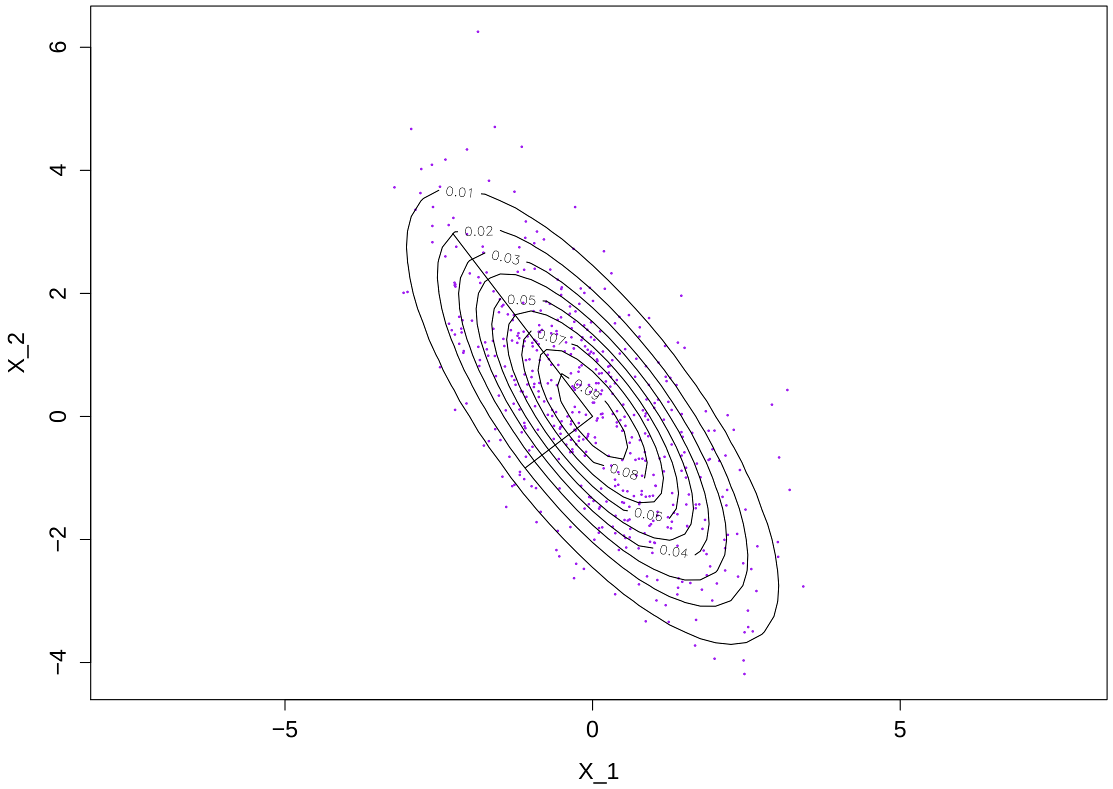

Figure 2: Non-diagonal covariance matrix $\boldsymbol{\Sigma}$ from Exercise 1.

In the more general case where the covariance matrix is not necessarily diagonal, its inverse matrix can be diagonalized with eigenvalue decomposition, because it is a symmetric square matrix. If $\boldsymbol{\Sigma}$ is positive definite, there exist an orthonormal matrix $\mathbf{V}$ and diagonal matrix $\boldsymbol{\Lambda}$ s.t.

$$\boldsymbol{\Sigma}^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{\mathbf{-1}}\mathbf{V}^{\mathbf{T}}.$$

Hence, we get the equation of the ellipsoid

$$1 = \sum_{i=1}^{p} \left( \frac{y_i}{\sqrt{b\lambda_i}} \right)^2 .$$

for the transformation

$$\mathbf{y} = (y_1, \ldots, y_p) = \mathbf{V}^T(\mathbf{x} - \boldsymbol{\mu}).$$

Now the length of the semiaxes are given by

$$\sqrt{b\lambda_1} \ \ldots, \sqrt{b\lambda_p},$$

where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of the covariance matrix $\mathbf{\Sigma}$ (diagonal of $\mathbf{\Lambda}$), and their directions are given by its eigenvectors (columns of $\mathbf{V}$). This is illustrated in Figure 2 for the normal distribution given in Exercise 1 (a).

**3.**

(a) The authors claim that although discriminative classifiers are traditionally considered superior compared to generative classifiers because of their lower asymptotic error (the error rate of the classifier as the sample size grows to infinity), generative classifiers converge faster to their asymptotic error rate, and thus may have a higher accuracy on small sample sizes.

(b) Given class labels $y$ and predictors $\mathbf{x} = (x_1, \ldots, x_p)$, the objective function that generative classifier $h_{\text{Gen}}$ maximizes (with respect to parameter vector $\boldsymbol{\beta}$) is the joint likelihood $p(\mathbf{x}, y)$ (or equivalently its logarithm $\log p(\mathbf{x}, y)$), while discriminative classifiers $h_{\text{Dis}}$ maximize directly the conditional likelihood $p(y|\mathbf{x})$ (or equivalently its logarithm $\log p(y|\mathbf{x})$) or 0-1 loss.

Two models that the authors discuss are the case of continuous predictors, where each $p(x_i|y)$ is normal distribution, and a discrete predictor case, where each $p(x_i|y)$ is a Bernoulli distribution. In both of the cases the predictors are assumed independent. In the first case the generative-discriminative pair is normal discriminant analysis (author seem to refer to QDA with diagonal covariance matrix) and logistic regression, and in the second case the pair is Naive Bayes and logistic regression.

(c) It seems that in most of the data sets the error rate of the generative classifiers (Naive Bayes and normal discriminant analysis) does indeed initially decrease faster than the error rate of the logistic regression as the sample size grows, but logistic regression has a smaller error rate with higher sample sizes. However, with the smaller data sets the logistic regression does not catch up generative classifiers, because the sample size cannot be grown high enough to reach its asymptotic error rate. As suggested in the introduction, although discriminative classifiers have better asymptotic performance, generative classifiers may outperform them on smaller sample sizes.