582631 Introduction to Machine Learning, Fall 2016
Exercise set 3 or: How I Learned to Stop Worrying and Love the Normal Distribution

Due November 24th–25th. NB: Deadline for returning solutions my email (in case you can't attend a session) is 12:15 on Friday. There may be changes in the exercise sessions — we'll keep you informed.

Problem 1 (3 + 3 + 3 + 3 points)

(a) (3 points) Recall that the p-dimensional multivariate Gaussian distribution is defined by a mean vector $\boldsymbol{\mu}$ and a covariance matrix Σ. If $\mathbf{X}$ is normal distributed with parameters $\boldsymbol{\mu}$ and Σ, then Σ contains the covariance of each pair of components of $\mathbf{X}$, i.e.,

$$\mathrm{Cov}(X_r, X_s) = E[(X_r - \mu_r)(X_s - \mu_s)] = \Sigma_{rs} = \Sigma_{sr},$$

for all $1 \leq r, s \leq p$. The diagonal terms Σ_{rr} are called variances. Recall further that the correlation coefficient is defined as

$$\mathrm{Cor}(X_r, X_s) = \frac{\mathrm{Cov}(X_r, X_s)}{\sqrt{\mathrm{Cov}(X_r, X_r)\,\mathrm{Cov}(X_s, X_s)}}$$

Consider the bivariate case $p = 2$. Let the variance of X_1 be 2.0 and the variance of X_2 be 3.0, and let both variables have mean zero, $\boldsymbol{\mu} = (0, 0)$. Find Σ such that $\mathrm{Cor}(X_1, X_2) = -0.75$. Draw $n = 200$ data points from the normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with the obtained parameters, and evaluate the empirical covariance matrix, $\hat{\Sigma}$, and the empirical correlation between X_1 and X_2.

Hint: The R functions `mvrnorm` (from library `MASS`), `cov`, and `cor` should do the job. You should observe that the empirical and exact values are somewhat close but not exactly the same.

(b) (3 points) Create a scatter plot of the $n = 200$ points you sampled. Also use the function `kde2d` to obtain an estimate of the data density and visualize the density using functions such as `contour`, `image`, and `persp`.

Hint: Each of the last three functions can take the output of `kde2d` directly as their argument. Study the scatter plot and the visualizations, and try to get a feeling on how they reflect the parameters $\boldsymbol{\mu}$ and Σ. Try changing the parameters and repeat to see the effect. You can also increase or decrease the sample size.

(c) (3 points) Next, generate an evenly spaced grid of points of the form

$$\mathbf{x} = (x_1, x_2) \in \{\, (i\delta, j\delta)\,;\, i, j \in \{\, -20, -19, \ldots, 19, 20 \,\}\,\}$$

with $\delta = 0.25$. In other words, the points evenly cover a square area from -5 to 5 along each axis with $41 \cdot 41 = 1681$ uniformly spaced points.

Evaluate the density $\mathcal{N}(\mathbf{x}\,;\, \boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}$ and Σ are the same as in item (a), at each grid point using the formula given in the lecture slides (Lecture 6, p. 12), and store the resulting values in a 41×41 matrix. Use `contour`, `image`, and `persp` again to visualize the density.

Hint: You can generate the grid by `expand.grid(.25*(-20:20),.25*(-20:20))`. This will produce a 1681×2 array with the grid points (x_1, x_2) as rows. The inverse of a matrix can be obtained by `solve`. Apply the density formula to each of them to obtain 1681 density values. (The first of them should be about 1.1307×10^{-19}.) To organize them into a square matrix, use `matrix(..., nrow=41)`. This matrix will be the argument of the three visualization functions.

(Exercises continued on the next page...)

(d) (3 points) Here comes the challenge. (But don't give up: you're almost there! You should really consider working together with each other to solve hard exercises like this one.)

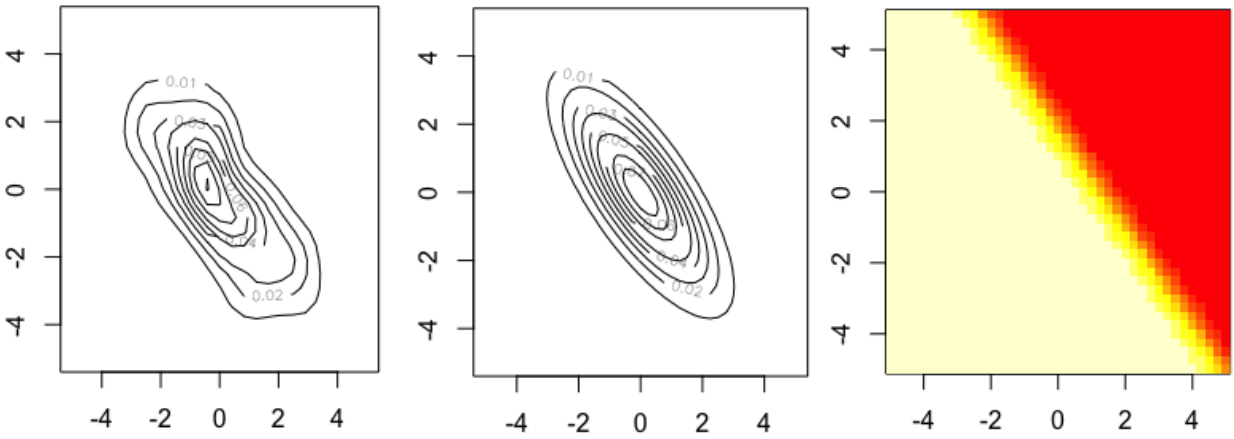
Denote the mean vector in items (a)–(c) by $\boldsymbol{\mu}_1$, and let $\boldsymbol{\mu}_2 = (2, 1)$. Compute the density at the same set of grid points as in item (c) under distribution $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma)$, i.e., with a different mean but the same covariance matrix.

Denote the two densities by $f_i(\mathbf{x}) = \mathcal{N}(\mathbf{x}\,;\, \boldsymbol{\mu}_i, \Sigma)$, $i \in \{\,1, 2\,\}$. Calculate the ratio

$$p(Y = 1 \mid \mathbf{x}) = \frac{f_1(\mathbf{x})\pi_1}{f_1(\mathbf{x})\pi_1 + f_2(\mathbf{x})\pi_2},$$

with $\pi_1 = \pi_2 = 1/2$. This is a linear discriminant (with uniform class distribution). Or to be more precise, this is the posterior probability of class $Y = 1$ given $\mathbf{x}$.

Visualize the decision boundary using, e.g., `contour`. As the name suggests, you should get a linear boundary. If you like, you can now try how well your classifier works by drawing data from either class and evaluating the above formula. What happens if you use different covariance matrices Σ_1 and Σ_2?



Here are our plots from `contour` *(b)–(c) and (for some variety)* `image` *(d).*

Problem 2 (4 + 2 points)

(a) (4 points) Prove the last equality on p. 11 of the slides (Lecture 6), i.e.,

$$\frac{1}{(2\pi)^{p/2}\sigma_1 \dots \sigma_p} \exp\left(-\frac{1}{2}\sum_{j=1}^{p}\frac{(x_j - \mu_j)^2}{\sigma_j^2}\right) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right),$$

where Σ is defined as on the same slide, $|\Sigma|$ is the determinant of Σ, and Σ^{-1} is the inverse of Σ.

Hint: The fact that Σ is a diagonal matrix makes your life a whole lot easier! Feel free to use Google to look for help.

(b) (2 points) See slide 13 (Lecture 6). Given $\mathcal{N}(\mathbf{x}\,;\, \boldsymbol{\mu}, \Sigma) = a$, solve for the value of b as a function of a such that

$$(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) = b.$$

Let $p = 3$, and Σ a diagonal matrix with $\Sigma_{jj} = j$ for all $j \in \{\,1, 2, 3\,\}$. Plug in the value $a = 1/100$ and and check that you get $b = 1.90495$.

(Exercises continued on the next page...)

Problem 3 (2+2+2 points)

Here you'll get a chance to get a taste of machine learning research through reading a scientific article, or as we call them, a paper. Download the paper "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes" by Andrew Ng and Mike Jordan (2001).[1]

(a) (2 points) Read the abstract and the Introduction. According to the authors, is discriminative learning better than generative learning? Justify your answer.

(b) (2 points) By a "parametric family of probabilistic models", the authors mean a set of distributions, where each distribution is defined by a set of parameters. An example of such a family is our friend, the family of normal distributions where the parameters are $\boldsymbol{\mu}$ and Σ. Ng and Jordan denote by h_{Gen} and h_{Dis} two models chosen by optimizing different things. What are these 'things' being optimized, i.e., what characterizes these two models? Which two families do the authors discuss, and what are the $(h_{\text{Gen}}, h_{\text{Dis}})$ pairs for those models?

(c) (2 points) Study Figure 1 in the paper. Explain what it suggests (see the last paragraph of the Introduction). Reflect this on item (a).

[1] `http://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf`