تمرین شماره ۱

تئوری بیز

تمرین های انتخاب شده از فصل ۲ کتاب

مهلت تحویل: ۹۹/۱/۱۰

شماره تمرین‌های	
۲	
۳	
۷	حل با متلب
۸	حل با متلب
۱۲	حل با متلب
۱۹	
۲۰	
۲۴	
۲۹	

where the definitions of $\phi_i(\cdot)$, $i = 1, 2, 3$, are readily understood by inspection. Underbraces indicate what variable the result of each summation depends on. To obtain $\phi_3(u)$ for each value of u, one needs to perform L operations (products and summations). Hence, a total number of L^2 operations is needed to compute $\phi_3(u)$ for all possible values of u. This is also true for the $\phi_2(u)$, $\phi_1(v)$. Thus, the total number of operations required to compute (2.141) is, after the factorization, of the order of L^2, instead of the order of L^5 demanded for the brute-force computation in (2.139). This procedure could be viewed as an effort to decompose a "global" sum into products of "local" sums to make computations tractable. Each summation can be viewed as a processing stage that removes a variable and provides as output a function. The essence of the algorithm given in [Li 94] is to search for the factorization that requires the minimal number of operations. This algorithm also has linear complexity in the number of nodes for singly connected networks. In general, for multiply connected networks the probability inference problem is NP-hard [Coop 90]. In light of this result, one tries to seek approximate solutions, as in [Dagu 93].

Training: Training of a Bayesian network consists of two parts. The first is to learn the network topology. The topology can either be fixed by an expert who can provide knowledge about dependencies or by use of optimization techniques based on the training set. Once the topology has been fixed, the unknown parameters (i.e., conditional probabilities and marginal probabilities) are estimated from the available training data points. For example, the fraction (frequency) of the number of instances that an event occurs over the total number of trials performed is a way to approximate probabilities. In Bayesian networks, other refined techniques are usually encountered. A review of learning procedures can be found in [Heck 95]. For the reader who wishes to delve further into the exciting world of Bayesian networks, the books of [Pear 88, Neap 04, Jens 01] will prove indispensable tools.

2.8 PROBLEMS

2.1 Show that in a multiclass classification task, the Bayes decision rule minimizes the error probability.

Hint: It is easier to work with the probability of correct decision.

2.2 In a two-class one-dimensional problem, the pdfs are the Gaussians $\mathcal{N}(0, \sigma^2)$ and $\mathcal{N}(1, \sigma^2)$ for the two classes, respectively. Show that the threshold x_0 minimizing the average risk is equal to

$$x_0 = 1/2 - \sigma^2 \ln \frac{\lambda_{21} P(\omega_2)}{\lambda_{12} P(\omega_1)}$$

where $\lambda_{11} = \lambda_{22} = 0$ has been assumed.

2.3 Consider a two equiprobable class problem with a loss matrix L. Show that if ϵ_1 is the probability of error corresponding to feature vectors from class ω_1

and ϵ_2 for those from class ω_2, then the average risk r is given by

$$r = P(\omega_1)\lambda_{11} + P(\omega_2)\lambda_{22} + P(\omega_1)(\lambda_{12} - \lambda_{11})\epsilon_1 + P(\omega_2)(\lambda_{21} - \lambda_{22})\epsilon_2$$

2.4 Show that in a multiclass problem with M classes the probability of classification error for the optimum classifier is bounded by

$$P_e \leq \frac{M-1}{M}$$

Hint: Show first that for each $\boldsymbol{x}$ the maximum of $P(\omega_i|\boldsymbol{x})$, $i = 1, 2, \ldots, M$, is greater than or equal to $1/M$. Equality holds if all $P(\omega_i|\boldsymbol{x})$ are equal.

2.5 Consider a two (equiprobable) class, one-dimensional problem with samples distributed according to the Rayleigh pdf in each class, that is,

$$p(x|\omega_i) = \begin{cases} \frac{x}{\sigma_i^2} \exp\left(\frac{-x^2}{2\sigma_i^2}\right) & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Compute the decision boundary point $g(x) = 0$.

2.6 In a two-class classification task, we constrain the error probability for one of the classes to be fixed, that is, $\epsilon_1 = \epsilon$. Then show that minimizing the error probability of the other class results in the likelihood test

$$\text{decide } \boldsymbol{x} \text{ in } \omega_1 \text{ if } \frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})} > \theta$$

where θ is chosen so that the constraint is fulfilled. This is known as the *Neyman–Pearson test*, and it is similar to the Bayesian minimum risk rule.

Hint: Use a Lagrange multiplier to show that this problem is equivalent to minimizing the quantity

$$q = \theta(\epsilon_1 - \epsilon) + \epsilon_2$$

2.7 In a three-class, two-dimensional problem the feature vectors in each class are normally distributed with covariance matrix

$$\Sigma = \begin{bmatrix} 1.2 & 0.4 \\ 0.4 & 1.8 \end{bmatrix}$$

The mean vectors for each class are $[0.1, 0.1]^T$, $[2.1, 1.9]^T$, $[-1.5, 2.0]^T$. Assuming that the classes are equiprobable, (a) classify the feature vector $[1.6, 1.5]^T$ according to the Bayes minimum error probability classifier; (b) draw the curves of equal Mahalanobis distance from $[2.1, 1.9]^T$.

2.8 In a two-class, three-dimensional classification problem, the feature vectors in each class are normally distributed with covariance matrix

$$\Sigma = \begin{bmatrix} 0.3 & 0.1 & 0.1 \\ 0.1 & 0.3 & -0.1 \\ 0.1 & -0.1 & 0.3 \end{bmatrix}$$

The respective mean vectors are $[0, 0, 0]^T$ and $[0.5, 0.5, 0.5]^T$. Derive the corresponding linear discriminant functions and the equation describing the decision surface.

2.9 In a two equiprobable class classification problem, the feature vectors in each class are normally distributed with covariance matrix Σ, and the corresponding mean vectors are $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$. Show that for the Bayesian minimum error classifier, the error probability is given by

$$P_B = \int_{(1/2)d_m}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)\, dz$$

where d_m is the Mahalanobis distance between the mean vectors. Observe that this is a decreasing function of d_m.
Hint: Compute the log-likelihood ratio $u = \ln p(\boldsymbol{x}|\omega_1) - \ln p(\boldsymbol{x}|\omega_2)$. Observe that u is also a random variable normally distributed as $\mathcal{N}((1/2)d_m^2, d_m^2)$ if $\boldsymbol{x} \in \omega_1$ and as $\mathcal{N}(-(1/2)d_m^2, d_m^2)$ if $\boldsymbol{x} \in \omega_2$. Use this information to compute the error probability.

2.10 Show that in the case in which the feature vectors follow Gaussian pdfs, the likelihood ratio test in (2.20)

$$\boldsymbol{x} \in \omega_1(\omega_2) \quad \text{if} \quad l_{12} \equiv \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} > (<)\theta$$

is equivalent to

$$d_m^2(\boldsymbol{\mu}_1, \boldsymbol{x}|\Sigma_1) - d_m^2(\boldsymbol{\mu}_2, \boldsymbol{x}|\Sigma_2) + \ln \frac{|\Sigma_1|}{|\Sigma_2|} < (>) - 2\ln\theta$$

where $d_m(\boldsymbol{\mu}_i, \boldsymbol{x}|\Sigma_i)$ is the Mahalanobis distance between $\boldsymbol{\mu}_i$ and $\boldsymbol{x}$ with respect to the Σ_i^{-1} norm.

2.11 If $\Sigma_1 = \Sigma_2 = \Sigma$, show that the criterion of the previous problem becomes

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} \boldsymbol{x} > (<) \Theta$$

where

$$\Theta = \ln\theta + 1/2(\|\boldsymbol{\mu}_1\|_{\Sigma^{-1}} - \|\boldsymbol{\mu}_2\|_{\Sigma^{-1}})$$

2.12 Consider a two-class, two-dimensional classification task, where the feature vectors in each of the classes ω_1, ω_2 are distributed according to

$$p(\boldsymbol{x}|\omega_1) = \frac{1}{\left(\sqrt{2\pi\sigma_1^2}\right)^2} \exp\left(-\frac{1}{2\sigma_1^2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T(\boldsymbol{x} - \boldsymbol{\mu}_1)\right)$$

$$p(\boldsymbol{x}|\omega_2) = \frac{1}{\left(\sqrt{2\pi\sigma_2^2}\right)^2} \exp\left(-\frac{1}{2\sigma_2^2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^T(\boldsymbol{x} - \boldsymbol{\mu}_2)\right)$$

with

$$\boldsymbol{\mu}_1 = [1, 1]^T, \quad \boldsymbol{\mu}_2 = [1.5, 1.5]^T, \quad \sigma_1^2 = \sigma_2^2 = 0.2$$

Assume that $P(\omega_1) = P(\omega_2)$ and design a Bayesian classifier

(a) that minimizes the error probability

(b) that minimizes the average risk with loss matrix

$$\Lambda = \begin{bmatrix} 0 & 1 \\ 0.5 & 0 \end{bmatrix}$$

Using a pseudorandom number generator, produce 100 feature vectors from each class, according to the preceding pdfs. Use the classifiers designed to classify the generated vectors. What is the percentage error for each case? Repeat the experiments for $\boldsymbol{\mu}_2 = [3.0, 3.0]^T$.

2.13 Repeat the preceding experiment if the feature vectors are distributed according to

$$p(\boldsymbol{x}|\omega_i) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\Big)$$

with

$$\Sigma = \begin{bmatrix} 1.01 & 0.2 \\ 0.2 & 1.01 \end{bmatrix}$$

and $\boldsymbol{\mu}_1 = [1, 1]^T, \boldsymbol{\mu}_2 = [1.5, 1.5]^T$.

Hint: To generate the vectors, recall from [Papo 91, p. 144] that a linear transformation of Gaussian random vectors also results in Gaussian vectors. Note also that

$$\begin{bmatrix} 1.01 & 0.2 \\ 0.2 & 1.01 \end{bmatrix} = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}$$

2.14 Consider a two-class problem with normally distributed vectors with the same Σ in both classes. Show that the decision hyperplane at the point $\boldsymbol{x}_0$, Eq. (2.46), is tangent to the constant Mahalanobis distance hyperellipsoids.

Hint: (a) Compute the gradient of Mahalanobis distance with respect to $\boldsymbol{x}$. (b) Recall from vector analysis that $\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}$ is normal to the tangent of the surface $f(\boldsymbol{x}) =$ constant.

2.15 Consider a two-class, one-dimensional problem with $p(x|\omega_1)$ being $\mathcal{N}(\mu, \sigma^2)$ and $p(x|\omega_2)$ a uniform distribution between a and b. Show that the Bayesian error probability is bounded by $G\left(\frac{b-\mu}{\sigma}\right) - G\left(\frac{a-\mu}{\sigma}\right)$, where $G(x) \equiv P(y \leq x)$ and y is $\mathcal{N}(0, 1)$.

2.16 Show that the mean value of the random vector $\frac{\partial \ln(p(\boldsymbol{x};\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}$ is zero.

2.17 In a heads or tails coin-tossing experiment the probability of occurrence of a head (1) is q and that of a tail (0) is $1 - q$. Let $x_i, i = 1, 2, \ldots, N$, be the resulting experimental outcomes, $x_i \in \{0, 1\}$. Show that the ML estimate of q is

$$q_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Hint: The likelihood function is

$$P(X : q) = \prod_{i=1}^{N} q^{x_i}(1-q)^{(1-x_i)}$$

Then show that the ML results from the solution of the equation

$$q^{\sum_i x_i}(1-q)^{(N-\sum_i x_i)}\left(\frac{\sum_i x_i}{q} - \frac{N - \sum_i x_i}{1-q}\right) = 0$$

2.18 The random variable x is normally distributed $\mathcal{N}(\mu, \sigma^2)$, where μ is considered unknown. Given N measurements of the variable, compute the Cramer–Rao bound $-E\left[\frac{\partial^2 L(\mu)}{\partial^2 \mu}\right]$ (Appendix A). Compare the bound with the variance of the resulting ML estimate of μ. Repeat this if the unknown parameter is the variance σ^2. Comment on the results.

2.19 Show that if the likelihood function is Gaussian with unknowns the mean $\boldsymbol{\mu}$ as well as the covariance matrix Σ, then the ML estimates are given by

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_k$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{k=1}^{N}(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}})^T$$

2.20 Prove that the covariance estimate

$$\hat{\Sigma} = \frac{1}{N-1}\sum_{k=1}^{N}(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}})^T$$

is an unbiased one, where

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_k$$

2.21 Prove that the ML estimates of the mean value and the covariance matrix (Problem 2.19) can be computed recursively, that is,

$$\hat{\boldsymbol{\mu}}_{N+1} = \hat{\boldsymbol{\mu}}_N + \frac{1}{N+1}(\boldsymbol{x}_{N+1} - \hat{\boldsymbol{\mu}}_N)$$

and

$$\hat{\Sigma}_{N+1} = \frac{N}{N+1}\hat{\Sigma}_N + \frac{N}{(N+1)^2}(\boldsymbol{x}_{N+1} - \hat{\boldsymbol{\mu}}_N)(\boldsymbol{x}_{N+1} - \hat{\boldsymbol{\mu}}_N)^T$$

where the subscript in the notation of the estimates, $\hat{\boldsymbol{\mu}}_N$, $\hat{\Sigma}_N$ indicates the number of samples used for their computation.

2.22 The random variable x follows the Erlang pdf

$$p(x;\theta) = \theta^2 x \exp(-\theta x)u(x)$$

where $u(x)$ is the unit-step function,

$$u(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Show that the maximum likelihood estimate of θ, given N measurements, $x_1, \ldots, x_N$, of x, is

$$\hat{\theta}_{ML} = \frac{2N}{\sum_{k=1}^{N} x_k}$$

2.23 In the ML estimation, the zero of the derivative of the log pdf derivative was computed. Using a multivariate Gaussian pdf, show that this corresponds to a maximum and not to a minimum.

2.24 Prove that the sum $z = x + y$ of two independent random variables, x and y, where $x \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, is also a Gaussian one with mean value and variance equal to $\mu_x + \mu_y$ and $\sigma_x^2 + \sigma_y^2$, respectively.

2.25 Show relations (2.74) and (2.75). Then show that $p(x|X)$ is also normal with mean μ_N and variance $\sigma^2 + \sigma_N^2$. Comment on the result.

2.26 Show that the posterior pdf estimate in the Bayesian inference task, for independent variables, can be computed recursively, that is,

$$p(\boldsymbol{\theta}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \frac{p(\boldsymbol{x}_N|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1})}{p(\boldsymbol{x}_N|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N-1})}$$

2.27 Show Eqs. (2.76)–(2.79).

2.28 The random variable x is normally distributed as $\mathcal{N}(\mu, \sigma^2)$, with μ being the unknown parameter described by the Rayleigh pdf

$$p(\mu) = \frac{\mu \exp(-\mu^2/2\sigma_\mu^2)}{\sigma_\mu^2}$$

Show that the maximum *a posteriori* probability estimate of μ is given by

$$\hat{\mu}_{MAP} = \frac{Z}{2R}\left(1 + \sqrt{1 + \frac{4R}{Z^2}}\right)$$

where

$$Z = \frac{1}{\sigma^2}\sum_{k=1}^{N} x_k, \quad R = \frac{N}{\sigma^2} + \frac{1}{\sigma_\mu^2}$$

2.29 Show that for the lognormal distribution

$$p(x) = \frac{1}{\sigma x\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \theta)^2}{2\sigma^2}\right), \quad x > 0$$

the ML estimate is given by

$$\hat{\theta}_{ML} = \frac{1}{N}\sum_{k=1}^{N} \ln x_k$$

2.30 Show that if the mean value and the variance of a random variable are known, that is,

$$\mu = \int_{-\infty}^{+\infty} xp(x)\,dx, \quad \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x)\,dx$$

the maximum entropy estimate of the pdf is the Gaussian $\mathcal{N}(\mu, \sigma^2)$.

2.31 Show Eqs. (2.98), (2.99), and (2.100).
Hint: For the latter, note that the probabilities add to one; thus a Lagrangian multiplier must be used.

2.32 Let P be the probability of a random point x being located in a certain interval h. Given N of these points, the probability of having k of them inside h is given by the binomial distribution

$$\text{prob}\{k\} = \frac{N!}{k!(N-k)!}P^k(1-P)^{N-k}$$

Show that $E[k/N] = P$ and that the variance around the mean is $\sigma^2 = E[(k/N - P)^2] = P(1-P)/N$. That is, the probability estimator $P = k/N$ is unbiased and asymptotically consistent.

2.33 Consider three Gaussian pdfs: $\mathcal{N}(1.0, 0.1), \mathcal{N}(3.0, 0.1)$, and $\mathcal{N}(2.0, 0.2)$. Generate 500 samples according to the following rule. The first two samples are generated from the second Gaussian, the third sample from the first one, and the fourth sample from the last Gaussian. This rule repeats until all 500 samples have been generated. The pdf underlying the random samples is modeled as a mixture

$$\sum_{i=1}^{3} \mathcal{N}(\mu_i, \sigma_i^2)P_i$$

Use the EM algorithm and the generated samples to estimate the unknown parameters μ_i, σ_i^2, P_i.

2.34 Consider two classes ω_1, ω_2 in the two-dimensional space. The data from class ω_1 are uniformly distributed inside a circle of radius r. The data of class ω_2 are also uniformly distributed inside another circle of radius r. The distance between the centers of the circles is greater than $4r$. Let N be the number of the available training samples. Show that the probability of error of the NN classifier is always smaller than that of the kNN, for any $k \geq 3$.

2.35 Generate 50 feature vectors for each of the two classes of Problem 2.12, and use them as training points. In the sequel, generate 100 vectors from each class and classify them according to the NN and 3NN rules. Compute the classification error percentages.

2.36 The pdf of a random variable is given by

$$p(x) = \begin{cases} \frac{1}{2} & \text{for } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Use the Parzen window method to approximate it using as the kernel function the Gaussian $\mathcal{N}(0, 1)$. Choose the smoothing parameter to be (a) $h = 0.05$ and (b) $h = 0.2$. For each case, plot the approximation based on $N = 32, N = 256$, and $N = 5000$ points, which are generated from a pseudorandom generator according to $p(x)$.

2.37 Repeat the preceding problem by generating $N = 5000$ points and using k nearest neighbor estimation with $k = 32, 64$, and 256, respectively.

2.38 Show that the variance $\sigma_N^2(\boldsymbol{x})$ of the pdf estimate, given by Eq. (2.110), is upper bounded by:

$$\sigma_N^2(\boldsymbol{x}) \leq \frac{\sup(\phi) E[\hat{p}(\boldsymbol{x})]}{N h^l}$$

where $\sup(\cdot)$ is the supremum of the associated function. Observe that for large values of h the variance is small. On the other hand, we can make the variance small for small values of h, provided N tends to infinity and if, also, the product Nh^l tends to infinity.

2.39 Recall Equation (2.128)

$$p(\boldsymbol{x}) = p(x_1) \prod_{i=2}^{l} p(x_i | A_i)$$

Assume $l = 6$ and

$$p(x_6 | x_5, \ldots, x_1) = p(x_6 | x_5, x_1) \tag{2.142}$$

$$p(x_5 | x_4, \ldots, x_1) = p(x_5 | x_4, x_3) \tag{2.143}$$

$$p(x_4 | x_3, x_2, x_1) = p(x_4 | x_3, x_2, x_1) \tag{2.144}$$

$$p(x_3 | x_2, x_1) = p(x_3) \tag{2.145}$$

$$p(x_2|x_1) = p(x_2) \tag{2.146}$$

Write the respective sets A_i, $i = 1, 2, \ldots, 6$, and construct the corresponding DAG.

2.40 In the DAG defined in Figure 2.29, assume that the variable z is measured to be $z0$. Compute $P(x1|z0)$ and $P(w0|z0)$.

2.41 In the example associated with the tree-structured DAG of Figure 2.28, assume that the patient undergoes the medical test H_1 and that this turns out to be positive (True). Based on this test, compute the probability that the patient has developed cancer. In other words, compute the conditional probability $P(C = \text{True}|H_1 = \text{True})$.

MATLAB PROGRAMS AND EXERCISES

Computer Exercises

A number of MATLAB functions are provided, which will help the interested reader to experiment on some of the most important issues discussed in the present chapter. Needless to say that there may be other implementations of these functions. Short comments are also given along with the code. In addition, we have used the symbols m and S to denote the mean vector (given as a column vector) and the covariance matrix, respectively, instead of the symbols $\boldsymbol{\mu}$ and Σ, which are used in the text. In the following, unless otherwise stated, each class is represented by an integer in $\{1, \ldots, c\}$ where c is the number of classes.

2.1 *Gaussian generator.* Generate N l-dimensional vectors from a Gaussian distribution with mean m and covariance matrix S, using the *mvnrnd* MATLAB function.

Solution

Just type

```
mvnrnd(m,S,N)
```

2.2 *Gaussian function evaluation.* Write a MATLAB function that computes the value of the Gaussian distribution $\mathcal{N}(m, S)$, at a given vector x.

Solution

```
function z=comp_gauss_dens_val(m,S,x)
  [l,q]=size(m);  % l=dimensionality
  z=(1/((2*pi)^(l/2)*det(S)^0.5) )...
      *exp(-0.5*(x-m)'*inv(S)*(x-m));
```