PATTERN RECOGNITION
ASSIGNMENT 2.
Parametric Learning, Nonparametric Techniques.
Due: Oct. 20, 2019.

Do the following problems from the textbook. [1]

1. (10 marks) Problem 4, sect. 3.2, p. 141.
   Let $\boldsymbol{x}$ be a d-dimensional binary (0 or 1) vector with a multivariate Bernoulli distribution

$$P(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i},$$

   where $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)^t$ is an unknown parameter vector, θ_i being the probability that $x_i = 1$. Show that the maximum likelihood estimate for $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x}_k.$$

2. (20 marks) Problem 17, sect. 3.5, p. 145.
   The purpose of this problem is to derive the Bayesian classifier for the d-dimensional multivariate Bernoulli case. As usual, work with each class separately, interpreting $P(\mathbf{x}|\mathcal{D})$ to mean $P(\mathbf{x}|\mathcal{D}_i, \omega_i)$. Let the conditional probability for a given category be given by

$$P(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i},$$

   and let $D = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ be a set of n samples independently drawn according to this probability density.

   (a) If $\boldsymbol{s} = (s_1, ..., s_d)^t$ is the sum of the n samples, show that

$$P(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i}.$$

   (b) Assuming a uniform a priori distribution for $\boldsymbol{\theta}$ and using the identity

$$\int_0^1 \theta^m(1-\theta)^n d\theta = \frac{m!n!}{(m+n+1)!},$$

   show that

$$p(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{d} \frac{(n+1)!}{s_i!(n-s_i)!}\theta_i^{s_i}(1-\theta_i)^{n-s_i}.$$

   (c) Plot this density for the case $d = 1, n = 1$ and for the two resulting possibilities for s_1.

   (d) Integrate the product $P(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})$ over $\boldsymbol{\theta}$ to obtain the desired conditional probability

$$P(\boldsymbol{x}|\mathcal{D}) = \prod_{i=1}^{d} \left(\frac{s_i+1}{n+2}\right)^{x_i}\left(1-\frac{s_i+1}{n+2}\right)^{1-x_i}.$$

[1] Duda, Hart, and Stork, *Pattern Classification*, Wiley, 2nd edition, 2001

(e) If we think of obtaining $P(\boldsymbol{x}|\mathcal{D})$ by substituting an estimate $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in $P(\boldsymbol{x}|\boldsymbol{\theta})$, what is the effective Bayesian estimate for $\boldsymbol{\theta}$?

3. (20 marks) Problem 2, sect. 4.3, p. 201.
Consider a normal $p(x) \sim N(\mu, \sigma^2)$ and Parzen-window function $\phi(x) \sim N(0, 1)$. Show that the Parzen-window estimate

$$p_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \phi\left(\frac{x - x_i}{h_n}\right)$$

has the following properties:

(a)

$$\overline{p}_n(x) \sim N(\mu, \sigma^2 + h_n^2)$$

(b)

$$Var[p_n(x)] \simeq \frac{1}{2nh_n\sqrt{\pi}} p(x)$$

(c)

$$p(x) - \overline{p}_n(x) \simeq \frac{1}{2}\left(\frac{h_n}{\sigma}\right)^2 \left[1 - \left(\frac{x-\mu}{\sigma}\right)^2\right] p(x)$$

for small h_n. (Note: if $h_n = h_1/\sqrt{n}$, this shows that the error due to bias goes to zero as $1/n$, whereas the standard deviation of the noise only goes to zero as $\sqrt[4]{n}$.)

4. (20 marks) Problem 8, sect. 4.5, p. 202.
It is easy to see that the nearest-neighbor error rate P can equal the Bayes rate P^*. if $P^* = 0$ (the best possibility) or if $P^* = (c-1)/c$ (the worst possibility). One might ask whether or not there are problems for which $P = P^*$ when P^* is between these extremes.

(a) Show that the Bayes rate for the one-dimensional case where $P(\omega_i) = 1/c$ and

$$P(x|\omega_i) = \begin{cases} 1 & 0 \le x \le \frac{cr}{c-1} \\ 1 & i \le x \le i + 1 - \frac{cr}{c-1} \\ 0 & \text{elsewhere} \end{cases}$$

is $P^* = r$.

(b) Show that for this case that the nearest-neighbor rate is $P = P^*$.

5. (30 marks) Problem 9, sect. 3.8, p. 158 (computer exercise).
Consider the Fischer linear discriminant method.

(a) Write a general program to calculate optimal direction $\mathbf{w}$ for a Fischer linear discriminant method based on three-dimensional data.

(b) Find the optimal $\mathbf{w}$ for categories ω_2 and ω_3 in the table

	ω_1			ω_2			ω_3		
sample	x_1	x_2	x_3	x_1	x_2	x_3	x_1	x_2	x_3
1	0.42	−0.087	0.58	−0.4	0.58	0.089	0.83	1.6	−0.014
2	−0.2	−3.3	−3.4	−0.31	0.27	−0.04	1.1	1.6	0.48
3	1.3	−0.32	1.7	0.38	0.055	−0.035	−0.44	−0.41	0.32
4	0.39	0.71	0.23	−0.15	0.53	0.011	0.047	−0.45	1.4
5	−1.6	−5.3	−0.15	−0.35	0.47	0.034	0.28	0.35	3.1
6	−0.029	0.89	−4.7	0.17	0.69	0.1	−0.39	−0.48	0.11
7	−0.23	1.9	2.2	−0.011	0.55	−0.18	0.34	−0.079	0.14
8	0.27	−0.3	−0.87	−0.27	0.61	0.12	−0.3	−0.22	2.2
9	−1.9	0.76	−2.1	−0.065	0.49	0.0012	1.1	1.2	−0.46
10	0.87	−1.0	−2.6	−0.12	0.054	−0.063	0.18	−0.11	−0.49

(c) Plot a line representing your optimal direction $\mathbf{w}$ and mark on it the positions of the projected points.

(d) In the subspace, fit each distribution with a (univariate) Gaussian, and find the resulting decision boundary.

(e) What is the training error (the error on the training points themselves) in the optimal subspace you found in (b)?

(f) For comparison, repeat parts (d) and (e) using instead the nonoptimal direction $\mathbf{w} = (1.0, 2.0, -1.5)^t$. What is the training error in this nonoptimal subspace?