Mid-term solutions
ECE 271A
Electrical and Computer Engineering
University of California San Diego

Nuno Vasconcelos Fall 2019

1. a) According to bayes decesion rule, the optimal decision function, under the "0/1" loss is "pick $i = 0$" if

$$\begin{aligned} P_{Y|X}(0|x) &> P_{Y|X}(1|x) \\ P_{X|Y}(x|0)P_Y(0) &> P_{X|Y}(x|1)P_Y(1) \end{aligned}$$

Taking Log on both sides,

$$\begin{aligned} \log \frac{P_{X|Y}(x|0)}{P_{X|Y}(x|1)} + \log \frac{P_Y(0)}{P_Y(1)} &> 0 \\ -\log \frac{\Gamma(k_0)}{\Gamma(k_1)} - k_0 \log b_0 + k_1 \log b_1 + (k_0 - k_1)\log x - x(\frac{1}{b_0} - \frac{1}{b_1}) + \log \frac{P_Y(0)}{P_Y(1)} &> 0 \end{aligned}$$

or

$$(k_0 - k_1)\log x - x(\frac{1}{b_0} - \frac{1}{b_1}) > T$$

with

$$T = \log \frac{\Gamma(k_0)}{\Gamma(k_1)} + k_0 \log b_0 - k_1 \log b_1 - \log \frac{P_Y(0)}{P_Y(1)}$$

b) The optimal decision rule is a simple threshold on x when (i) $k_0 = k_1$ and (ii) $b_0 = b_1$.

When $k_0 = k_1 = k$ and assuming $b_0 > b_1$,

$$x > \frac{b_1 b_0}{b_0 - b_1}\left(k \log \frac{b_0}{b_1} - \log \frac{P_Y(0)}{P_Y(1)}\right)$$

When $b_0 = b_1 = b$ and assuming $k_0 > k_1$,

$$\begin{aligned} (k_0 - k_1)(\log x - \log b) &> \log \frac{\Gamma(k_0)}{\Gamma(k_1)} - \log \frac{P_Y(0)}{P_Y(1)} \\ \log \frac{x}{b} &> \frac{1}{k_0 - k_1}\left(\log \frac{\Gamma(k_0)}{\Gamma(k_1)} - \log \frac{P_Y(0)}{P_Y(1)}\right) \\ x &> b e^{\frac{1}{k_0 - k_1}\left(\log \frac{\Gamma(k_0)}{\Gamma(k_1)} - \log \frac{P_Y(0)}{P_Y(1)}\right)} \end{aligned}$$

c) The ML estimate of the parameter b is given by,

$$b^* = \arg\max_b P_X(\mathcal{D})$$

where,

$$\begin{aligned} P_X(\mathcal{D}) &= \prod_1^n P_X(x_i) \\ \log P_X(\mathcal{D}) &= \sum_1^n \log P_X(x_i) \\ &= \sum_1^n -\log\Gamma(k) - k\log b + (k-1)\log x_i - \frac{x_i}{b} \\ &= -n\log\Gamma(k) - nk\log b + (k-1)\sum_1^n \log x_i - \frac{1}{b}\sum_1^n x_i \end{aligned}$$

Taking gradient with respect to b, and setting to zero

$$\frac{\partial \log P_X(\mathcal{D})}{\partial b} = -\frac{nk}{b} + \frac{1}{b^2}\sum_1^n x_i = 0$$

It follows that,

$$b^* = \frac{\sum_1^n x_i}{nk}$$

Computing the second derivative,

$$\frac{\partial^2 \log P_X(\mathcal{D})}{\partial b^2} = \frac{nk}{b^2} - 2\frac{\sum_1^n x_i}{b^3} = -\frac{nk}{b^{*2}}$$

is a negative quantity when $k > 0$, therefore we have a maximum at $b^* = \frac{\sum_1^n x_i}{nk}$.

d) If the ML extimate of **c)** is unbaised,

$$\begin{aligned} E(b^* - b)) &= 0 \\ E(\frac{\sum_1^n X_i}{nk} - b) &= 0 \\ \frac{\sum_1^n E(X)}{n} &= bk \\ E(X) &= bk \end{aligned}$$

2.a) The log-likelihood function is

$$\mathcal{L}(\mathcal{D}) = -T \log \sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2} \sum_t (x_t - \alpha - \beta t)^2. \tag{1}$$

The gradient with respect to (α, β) is zero when

$$\begin{aligned} \frac{\partial l(\mathcal{D})}{\partial \alpha} &= \frac{1}{\sigma^2} \sum_t (x_t - \alpha - \beta t) = 0 \\ \frac{\partial l(\mathcal{D})}{\partial \beta} &= \frac{1}{\sigma^2} \sum_t t(x_t - \alpha - \beta t) = 0 \end{aligned}$$

This can be written in matrix form as

$$\begin{bmatrix} T & \sum_t t \\ \sum_t t & \sum_t t^2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \sum_t x_t \\ \sum_t t x_t \end{bmatrix} \tag{2}$$

i.e.

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 1 & < t > \\ < t > & < t^2 > \end{bmatrix}^{-1} \begin{bmatrix} < x_t > \\ < t x_t > \end{bmatrix} \tag{3}$$

b) The Hessian is the matrix of derivatives

$$\begin{aligned} \frac{\partial^2 l(\mathcal{D})}{\partial \alpha^2} &= -\frac{T}{\sigma^2} \\ \frac{\partial^2 l(\mathcal{D})}{\partial \beta^2} &= -\frac{T}{\sigma^2} < t^2 > \\ \frac{\partial^2 l(\mathcal{D})}{\partial \alpha \beta} &= -\frac{T}{\sigma^2} < t > \end{aligned}$$

i.e.

$$H = -\frac{T}{\sigma^2} \begin{bmatrix} 1 & < t > \\ < t > & < t^2 > \end{bmatrix} \tag{4}$$

This has determinant $-\frac{T}{\sigma^2}(< t^2 > - < t >^2) = -\frac{T}{\sigma^2} \left\langle (t - < t >)^2 \right\rangle$ which is allways negative. Since the top left entry is also negative, the Hessian is negative definite and we have a maximum.

c) To compute the bias, we note that

$$\begin{aligned} E_{X_1,\ldots,X_T} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} &= \begin{bmatrix} 1 & < t > \\ < t > & < t^2 > \end{bmatrix}^{-1} E_{X_1,\ldots,X_T} \begin{bmatrix} \frac{1}{T} \sum_t x_t \\ \frac{1}{T} \sum_t t x_t \end{bmatrix} \\ &= \begin{bmatrix} 1 & < t > \\ < t > & < t^2 > \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{T} \sum_t \mu_t \\ \frac{1}{T} \sum_t t \mu_t \end{bmatrix} \\ &= \begin{bmatrix} 1 & < t > \\ < t > & < t^2 > \end{bmatrix}^{-1} \begin{bmatrix} a + b < t > \\ a < t > + b < t^2 > \end{bmatrix} \\ &= \begin{bmatrix} 1 & < t > \\ < t > & < t^2 > \end{bmatrix}^{-1} \begin{bmatrix} 1 & < t > \\ < t > & < t^2 > \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \\ &= \begin{bmatrix} a \\ b \end{bmatrix} \end{aligned}$$

from which the estimators are unbiased.

3. a)

$$\begin{aligned}
\mu_z &= E[z] = E[\mathbf{v}^T\mathbf{x}] = \mathbf{v}^T\mu \\
\sigma_z^2 &= E_z[(z-\mu_z)^2] = E_z[(\mathbf{v}^T(\mathbf{x}-\mu))^2] = E_z[\mathbf{v}^T(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\mathbf{v}] \\
&= \mathbf{v}^T E_z[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T]\mathbf{v} \\
&= \mathbf{v}^T\mathbf{\Sigma}\mathbf{v}
\end{aligned}$$

b) When

$$\mathbf{v} = \mathbf{w}_{ij}$$

(5) and (6) are the same. This is because z has class means

$$\begin{aligned}
\mu_{z,i} &= \mathbf{w}_{ij}^T\mu_i \\
&= (\mu_i-\mu_j)^T\mathbf{\Sigma}^{-1}\mu_i
\end{aligned}$$

and variance

$$\begin{aligned}
\sigma_z^2 &= \mathbf{w}_{ij}^T\mathbf{\Sigma}\mathbf{w}_{ij} \\
&= (\mu_i-\mu_j)^T\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}(\mu_i-\mu_j) \\
&= (\mu_i-\mu_j)^T\mathbf{\Sigma}^{-1}(\mu_i-\mu_j)
\end{aligned}$$

The optimal threshold on z is then

$$T_z = \frac{\mu_{z,i}+\mu_{z,j}}{2} - \frac{\sigma_z^2}{(\mu_{z,i}-\mu_{z,j})}\log\frac{P_Y(i)}{P_Y(j)}.$$

On the other hand

$$\begin{aligned}
\mathbf{w}_{ij}^T(\mathbf{x}-\mathbf{x}_{ij,0}) &= \mathbf{w}_{ij}^T\mathbf{x} - (\mu_i-\mu_j)^T\mathbf{\Sigma}^{-1}\left(\frac{\mu_i+\mu_j}{2} - \frac{\mu_i-\mu_j}{(\mu_i-\mu_j)^T\mathbf{\Sigma}^{-1}(\mu_i-\mu_j)}\log\frac{P_Y(i)}{P_Y(j)}\right) \\
&= z - \left(\frac{\mu_{z,i}+\mu_{z,j}}{2} - \frac{\sigma_z^2}{\mu_{z,i}-\mu_{z,j}}\log\frac{P_Y(i)}{P_Y(j)}\right) \\
&= z - T_z.
\end{aligned}$$

Course Outline
ECE271A – Statistical Learning I
Department of Electrical and Computer Engineering
University of California, San Diego
Nuno Vasconcelos

Your responsibilities in this class fall into three main categories:

1. Class participation and homework 20%
2. Mid-term 35%
3. Final 45%

You are allowed to collaborate on homework as long as you write your solutions independently and acknowledge the collaboration in the problems where it was used. Homework is due one week after the hand-out date. It will have a problem solving component and a component of computer problems. I assume that students have access to Matlab. The computer problems will consist of the application of a number of techniques to a given problem (Cheetah).

Instructor
Nuno Vasconcelos,
EBU1 5602, 4-5550, e-mail: nuno@ece.ucsd.edu
Office hours: Fridays 9:30-10:30AM

TA
See website.

Exam dates:

- Mid-term - TBA
- Final - finals week

Text: We will follow closely

- Richard O. Duda, Peter E. Hart and David G. Stork *Pattern Classification.* New York, NY: John Wiley&Sons, 2001.

Supplementary hand-outs will be distributed when appropriate. There are various other books of interest. These are not required but can be used for alternative explanations of the material.

1. C. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.
2. T. Hastie, R. Tibshirani, J. H. Friedman, *The Elements of Statistical Learning.* Springer Verlag, 2001.
3. Luc Devroye, Laszlo Gyorfi, Gabor Lugosi, *A Probabilistic Theory of Pattern Recognition.* Springer Verlag, 1998.
4. Andrew Gelman, Donald B. Rubin, Hal S. Stern, *Bayesian Data Analysis, Second Edition*, CRC Press; 2nd edition, 2003.

5. Tom Mitchell, *Machine Learning*, McGraw-Hill, 1997.
6. Christopher Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1996.
7. Vladimir Vapnik, *The Nature of Statistical Learning Theory*. Springer Verlag, 1999.

There is a web page for the course,

`http://www.svcl.ucsd.edu/courses/ece271A/ece271A.htm`

(also accessible from `http://www.svcl.ucsd.edu/~nuno`)

LECTURE SUBJECT	Number of classes
Introduction	1
Bayesian decision theory	2
The Gaussian classifier	1
Maximum likelihood estimation	1
Bias and variance	2
Bayesian parameter estimation	2
Conjugate and non-informative priors	1
Dimensionality and dimensionality reduction	2
The nearest neighbor classifier	1
Kernel-based density estimation	1
Mixture models and EM	3
Applications	1