Statistical Machine Learning 2016

Exercises and answers, week 6

6 October 2016

Exercise 1

Find the eigenvalues and a set of mutually orthogonal eigenvectors of the symmetric matrix:

$$\begin{bmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{bmatrix}$$

ANSWER: In order to find the eigenvalues of the matrix, which we will henceforth denote as A, we have to solve the characteristic equation ($\det(A - \lambda I) = 0$).

$$|A - \lambda I| = \begin{vmatrix} 3-\lambda & 2 & 4 \\ 2 & -\lambda & 2 \\ 4 & 2 & 3-\lambda \end{vmatrix} = -\lambda^3 + 6\lambda^2 + 15\lambda + 8 = -(\lambda - 8)(\lambda + 1)^2$$

The characteristic equation for our matrix is $(\lambda - 8)(\lambda + 1)^2 = 0$ and it has roots $\lambda_1 = \lambda_2 = -1$ and $\lambda_3 = 8$. Note that -1 is a double root. We now have to find two (orthogonal) eigenvectors for $\lambda = -1$ and one eigenvector for $\lambda = 8$.

First, let us solve $Av = -v$ corresponding to the eigenvalue $\lambda = -1$.

$$\begin{bmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = - \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \iff \begin{cases} 3v_1 + 2v_2 + 4v_3 = -v_1 \\ 2v_1 + 2v_3 = -v_2 \\ 4v_1 + 2v_2 + 3v_3 = -v_3 \end{cases} \iff \begin{cases} 4v_1 + 2v_2 + 4v_3 = 0 \\ 2v_1 + v_2 + 2v_3 = 0 \\ 4v_1 + 2v_2 + 4v_3 = 0 \end{cases}$$

We can see that this system reduces to the single equation $2v_1 + v_2 + 2v_3 = 0$. We have three variables to determine, but only one equation, so we arbitrarily choose for example $v_1 = s$ and $v_3 = t$ as parameters and use them to express v_2. Thus, the two eigenvectors of $\lambda = -1$ must have the form:

$$\begin{bmatrix} s \\ -2s - 2t \\ t \end{bmatrix}. \tag{1}$$

We now have to choose values for s and t that yield two orthogonal vectors. We can arbitrarily set the parameter values to $s = 1$ and $t = 0$ to get the first eigenvector (the only restriction is that it has to be a non-zero vector):

$$\mathbf{u}_1 = \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix} \quad \left(\text{Verify} : \begin{bmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix} = (-1) \cdot \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix} \right)$$

Now, we have to find a vector $\mathbf{u}_2$ of the form in Equation 1 such that $\mathbf{u}_1^T \mathbf{u}_2 = 0$.

$$\mathbf{u}_1^T \mathbf{u}_2 = 0 \iff s + (-2)(-2s - 2t) = 0 \iff 5s + 4t = 0 \tag{2}$$

We can choose for example $s = 4$ and $t = -5$, which satisfy Equation 2. We then get:

$$\mathbf{u}_2 = \begin{bmatrix} 4 \\ 2 \\ -5 \end{bmatrix} \quad \left(\text{Verify : } \begin{bmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{bmatrix} \begin{bmatrix} 4 \\ 2 \\ -5 \end{bmatrix} = \begin{bmatrix} -4 \\ -2 \\ 5 \end{bmatrix} = (-1) \cdot \begin{bmatrix} 4 \\ 2 \\ -5 \end{bmatrix} \right)$$

To get our final eigenvector, we have to solve $Av = 8v$, corresponding to the eigenvalue $\lambda = 8$.

$$\begin{bmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 8 \cdot \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \iff \begin{cases} 3v_1 + 2v_2 + 4v_3 = 8v_1 \\ 2v_1 + 2v_3 = 8v_2 \\ 4v_1 + 2v_2 + 3v_3 = 8v_3 \end{cases} \iff \begin{cases} -5v_1 + 2v_2 + 4v_3 = 0 \\ 2v_1 - 8v_2 + 2v_3 = 0 \\ 4v_1 + 2v_2 - 5v_3 = 0 \end{cases}$$

By solving the linear equation system above, we can show that the eigenvectors for $\lambda = 8$ are of the form:

$$\begin{bmatrix} 2r \\ r \\ 2r \end{bmatrix} \tag{3}$$

It is easy to check that this vector is orthogonal to $\mathbf{u}_1$ and $\mathbf{u}_2$ (and in general to all vectors of the form in Equation 1) for any choice of r, so let's take for example $r = 1$. We then get:

$$\mathbf{u}_3 = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} \quad \left(\text{Verify : } \begin{bmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 16 \\ 8 \\ 16 \end{bmatrix} = 8 \cdot \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} \right)$$

Note that since this real-valued matrix is symmetric we do indeed have 3 eigenvalues and a set of 3 orthogonal (and thus linearly independent) eigenvectors (one for each eigenvalue).

Exercise 2

Consider a discrete variable x that can take K values, $x \in \{1, \ldots, K\}$. If we denote the probability of $x = k$ by the parameter θ_k, then the distribution of x is given by

$$P(x = k|\boldsymbol{\theta}) = \theta_k \tag{4}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T$ and the parameters are constrained to satisfy

$$\theta_k \geq 0 \qquad \text{and} \qquad \sum_{k=1}^{K} \theta_k = 1 \tag{5}$$

1. Explain why the parameters should satisfy these constraints.

   ANSWER: Probabilities are non-negative, and should add up to one.

Now consider a dataset χ of N independent observations, $\chi = \{x_1, \ldots, x_N\}$.

2. Show that the log-likelihood $\ln P(\chi|\boldsymbol{\theta})$ is of the form

$$\ln P(\chi|\boldsymbol{\theta}) = \sum_{k=1}^{K} m_k \ln \theta_k \tag{6}$$

   What are the m_k's (in terms of the x_i's, k's etc.)?

   ANSWER: Note first that each of the data points x_n has a value in $1, \ldots, K$ and that $P(x_n|\boldsymbol{\theta}) = \theta_{x_n}$. So

$$\ln P(\chi|\boldsymbol{\theta}) = \sum_{n=1}^{N} \ln \theta_{x_n}.$$

Now it is convenient to introduce a Kronecker delta and rewrite and reshuffle a bit,

$$\begin{aligned}\ln P(\chi|\boldsymbol{\theta}) &= \sum_{n=1}^{N} \ln \theta_{x_n} \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} \delta_{x_n k} \ln \theta_k \\ &= \sum_{k=1}^{K}\sum_{n=1}^{N} \delta_{x_n k} \ln \theta_k \\ &= \sum_{k=1}^{K} m_k \ln \theta_k\end{aligned}$$

in which the 'counts' $m_k = \sum_{n=1}^{N} \delta_{x_n k}$ are the number of observations $x_n = k$.

3. Show that the maximum likelihood solution $\boldsymbol{\theta}^*$ is given by

$$\theta_k^* = \frac{m_k}{N} \tag{7}$$

Hint: Use a Lagrange multiplier for the constraint $\sum_{k=1}^{K} \theta_k - 1 = 0$.

ANSWER: Define the Lagrangian

$$L(\theta_1, \ldots, \theta_K, \lambda) = \sum_{k=1}^{K} m_k \ln \theta_k + \lambda(\sum_{k=1}^{K} \theta_k - 1)$$

Take the derivative with respect to θ_k and set to zero,

$$0 = \frac{\partial L}{\partial \theta_k} = \frac{m_k}{\theta_k} + \lambda$$

which gives

$$\theta_k = -\frac{m_k}{\lambda}$$

We solve for λ by substitution of the constraint $\sum_k \theta_k = 1$, so

$$\sum_{k=1}^{K} \theta_k = -\sum_{k=1}^{K} \frac{m_k}{\lambda} = -\frac{N}{\lambda} = 1$$

(note that $\sum_K m_k = N$), so we find $\lambda = -N$ and

$$\theta_k^* = \frac{m_k}{N}$$

Method 2: without Lagrange multipliers. Note that we can get rid of the constraints by considering log-likelihood as a function that directly depends on the $K-1$ independent parameters $\theta_1, \ldots, \theta_{K-1}$, and indirectly via the last dependent parameter θ_K, which is now considered as a function of the $K-1$ independent parameters

$$\theta_K(\theta_1, \ldots, \theta_{K-1}) = 1 - \sum_{k=1}^{K-1} \theta_k.$$

The log-likelihood as function of the independent parameters is then

$$L(\theta_1, \ldots, \theta_{K-1}) = \sum_{k=1}^{K-1} m_k \ln \theta_k + m_K \ln \theta_K(\theta_1, \ldots, \theta_{K-1})$$

Set partial derivatives to zero. Use the chain rule together with $\partial\theta_K/\partial\theta_k = -1$ for the last term,

$$0 = \frac{\partial L}{\partial \theta_k} = \frac{m_k}{\theta_k} - \frac{m_K}{\theta_K}, \quad \text{for } k = 1, \ldots, K-1$$

Multiply all partial derivatives by $\theta_k\theta_K$, then expand θ_K and substitute $m_K = N - \sum_{k=1}^{K-1} m_k$ (and use other letters for dummy indices) to obtain

$$\begin{aligned} 0 &= m_k\theta_K - m_K\theta_k \\ &= m_k(1 - \sum_{i=1}^{K-1} \theta_i) - (N - \sum_{i=1}^{K-1} m_i)\theta_k \quad \text{for } k = 1, \ldots, K-1 \end{aligned}$$

So note that this is a linear system, consisting of $K-1$ linear equations with $K-1$ unknowns (namely, $\theta_1, \ldots, \theta_{K-1}$.) We "add up the rows" by summing over k from which we obtain

$$\begin{aligned} 0 &= \sum_{k=1}^{K-1} m_k(1 - \sum_{i=1}^{K-1} \theta_i) - (N - \sum_{i=1}^{K-1} m_i) \sum_{k=1}^{K-1} \theta_k \\ &= \sum_{k=1}^{K-1} m_k - \sum_{k=1}^{K-1}\sum_{i=1}^{K-1} m_k\theta_i - N\sum_{k=1}^{K-1} \theta_k + \sum_{i=1}^{K-1}\sum_{k=1}^{K-1} m_i\theta_k \\ &= \sum_{k=1}^{K-1} m_k - N\sum_{k=1}^{K-1} \theta_k \end{aligned}$$

So

$$\sum_{k=1}^{K-1} \theta_k = \frac{\sum_{k=1}^{K-1} m_k}{N}$$

and therefore also

$$1 - \sum_{k=1}^{K-1} \theta_k = \frac{N - \sum_{k=1}^{K-1} m_k}{N}.$$

In other words,

$$\theta_K = \frac{m_K}{N}.$$

Now we plug this back in the linear system and find

$$\begin{aligned} 0 &= m_k\theta_K - m_K\theta_k \\ &= m_k\frac{m_K}{N} - m_K\theta_k \end{aligned}$$

from which we finally can conclude

$$\theta_k = \frac{m_k}{N}$$

Exercise 3

The beta distribution is

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \qquad 0 \leq \mu \leq 1 \tag{8}$$

in which $\Gamma(x)$ is the gamma function (a well defined mathematical function, see book, www exercise 1.17). The gamma function is a generalization of the factorial function $(n-1)!$ as it satisfies

$$\Gamma(x+1) = x\Gamma(x) \tag{9}$$

We are looking for an expression for the expectation value in terms of a and b

$$\langle\mu\rangle = \int_0^1 \mu \,\mathrm{Beta}(\mu|a,b)d\mu \tag{10}$$

Since the beta distribution is normalised, we can start from the relation

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{11}$$

1. Show that the expectation value is given by

$$\langle\mu\rangle = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \tag{12}$$

   Hint: you do not actually have to compute any integrals.

   ANSWER:

$$\begin{aligned}
\langle\mu\rangle &= \int_0^1 \mu\mathrm{Beta}(\mu|a,b)d\mu \\
&= \int_0^1 \mu \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}d\mu \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \mu^{a}(1-\mu)^{b-1}d\mu \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}
\end{aligned}$$

2. Use this result and the property $\Gamma(x+1) = x\Gamma(x)$ to show that

$$\langle\mu\rangle = \frac{a}{a+b} \tag{13}$$

   ANSWER: Continuing from 1:

$$\begin{aligned}
\langle\mu\rangle &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} \\
&= \frac{a}{a+b}
\end{aligned}$$

Exercise 4

Suppose we have two coins, A and B, and we do not know whether these coins are fair.

1. Let μ be the probability the coin comes up H(eads). Give an expression for the likelihood of a data set $\mathcal{D}$ of N observations of independent tosses of the coin.

   ANSWER: Obviously a Bernoulli distribution with $p(H) = p(x = 1) = \mu$, and so for the likelihood of a data set $\mathcal{D} = \{x_1, \ldots, x_N\}$ we have

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} = \mu^m (1-\mu)^l \tag{14}$$

   with m the number of observations of H(eads) in the data set and l the number of T(ails).

Suppose we have observed the following results of two series of coin tosses:

coin:	data $\mathcal{D}$:
A	H,T,T,H,T,T,T
B	H

2. What is the maximum likelihood estimate for μ_A, the probability that a toss with coin A results in H(eads)? And for μ_B? Based on these maximum likelihood estimates, what is the probability that the next toss of coin A will result in H(eads)? And the next toss with coin B? Do these results make sense?

   ANSWER: $\mu_{ML} = m/N$, so $\mu_{A,ML} = 2/7$ and $\mu_{B,ML} = 1/1$. This means that the maximum likelihood estimate would predict that the probability that the next toss of coin A will result in H is 2/7, and that with absolute certainty, the next toss of coin B will be H! Although the former seems to be more or less in agreement with common sense, the latter does not make sense at all; it is an example of severe overfitting of the ML solution.

3. Let us now take a Bayesian approach. Find an expression for $p(\mu|\mathcal{D})$ using Bayes' rule and show that a prior proportional to powers of μ and $(1-\mu)$ will lead to a posterior that is also proportional to powers of μ and $(1-\mu)$. Are you free to choose whatever prior you like?

   ANSWER: From Bayes' rule

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)\, p(\mu)}{p(\mathcal{D})} \propto p(\mathcal{D}|\mu)\, p(\mu) \tag{15}$$

   As $p(\mathcal{D}|\mu)$ in (14) consist of a product of powers of μ and $(1-\mu)$, multiplying by something proportional to the same will result in an expression (the posterior) that remains proportional to powers of μ and $(1-\mu)$.
   Note there is nothing priviliged about such a prior that would make it a 'better' choice than any other prior. A prior should, in principle, represent exactly what we know about μ *before* observing any data used to calculate the posterior. In many cases this makes it a pretty difficult problem: not only is it far from easy to convert all your prior knowledge into a correct probability distribution, but the subsequent computations can also become very hard (integrating over products of functions with complex structures). A reasonable alternative, albeit primarily a convenient one, is to choose a prior that captures important features of the 'true' prior quite well but has a functional form that makes it relatively easy to handle in combination with a certain likelihood: this is known as a *conjugate* prior.

Such a prior exists and is called the Beta distribution with hyperparameters a and b:

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \qquad 0 \le \mu \le 1 \tag{16}$$

in which $\Gamma(x)$ is the gamma function with property $\Gamma(x+1) = x\Gamma(x)$.

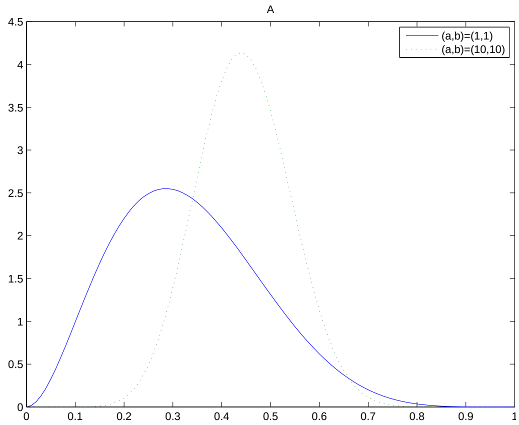
4. Give combinations (a, b) for a prior that expresses: a) total ignorance, b) high confidence in a reasonably fair coin. For each prior and each coin, calculate the posterior probability density of μ given the observed coin tosses $\mathcal{D}$ and plot the results (for example by using the betapdf command in MatLab). Do these results make more sense than the ML estimates?

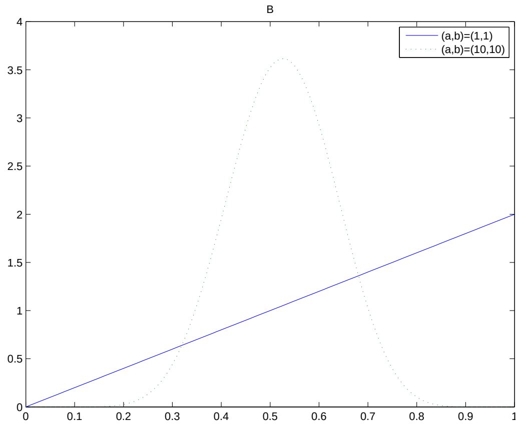
   ANSWER: We choose priors Beta(a, b) with, for example: a) $(1, 1)$, b) $(10, 10)$ (see also Bishop, Fig.2.2).
   First of all, from Bayes' rule:

$$\begin{aligned} p(\mu|\mathcal{D}, a, b) &\propto p(\mathcal{D}|\mu)\; p(\mu) \\ &\propto \mu^m (1-\mu)^l \; \mu^{a-1}(1-\mu)^{b-1} \\ &= \mu^{m+a-1}(1-\mu)^{l+b-1} \\ &\propto \text{Beta}(\mu|m+a, b+l) \end{aligned}$$

   The following MATLAB code will produce the requested plots:

```
X=[0:0.01:1];
m=2; l=5;
plot(X,betapdf(X,1+m,1+l),'-',X,betapdf(X,10+m,10+l),':');
legend('(a,b)=(1,1)','(a,b)=(10,10)');
title('A');
figure;
m=1; l=0;
plot(X,betapdf(X,1+m,1+l),'-',X,betapdf(X,10+m,10+l),':');
legend('(a,b)=(1,1)','(a,b)=(10,10)');
title('B');
```





Exercise 5

Consider a mixture of K Gaussian densities of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{17}$$

Show that if the mixing coefficients satisfy

$$\pi_k \geq 0 \quad \text{and} \quad \sum_k \pi_k = 1$$

then the mixture of Gaussians in (17) is positive and normalized. (You may assume that the components of the mixture are normalized)

ANSWER:
Positive: At least one of the π_k's is positive (due to normalization). $\mathcal{N}(x|\mu_k, \Sigma_k)$ is positive everywhere. So the product of the positive π_k and the Gaussian is positive everywhere. Other terms can not make it less positive.

Normalization: Integrate (17) over $\mathbf{x}$. Interchange sum and integral and then do the integrals over $\mathbf{x}$, so the Gaussians disappear (since they are normalized). Then a sum of mixture term remains, which sum to one.
In formulas:

$$\begin{aligned}\int p(\mathbf{x})d\mathbf{x} &= \int \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)d\mathbf{x} \\ &= \sum_{k=1}^{K} \int \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)d\mathbf{x} \\ &= \sum_{k=1}^{K} \pi_k \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)d\mathbf{x} \\ &= \sum_{k=1}^{K} \pi_k \cdot 1 = 1\end{aligned}$$

Conclusion: equation (17) does indeed represent a proper probability distribution.