Statistical Machine Learning 2016

Exercises and answers, week 2

8 September 2016

Exercise 1

(Exercise 1.14 from Bishop.)

1. Show that a matrix $\mathbf{W}$ with elements w_{ij} can be written as the sum of a symmetric matrix $\mathbf{W}^S$ and an anti-symmetric matrix $\mathbf{W}^A$. In other words, show that

$$w_{ij} = w_{ij}^S + w_{ij}^A \tag{1}$$

with symmetric matrix elements $w_{ij}^S = (w_{ij} + w_{ji})/2$ and anti-symmetric matrix elements $w_{ij}^A = (w_{ij} - w_{ji})/2$. Verify that $w_{ij}^S = w_{ji}^S$ and $w_{ij}^A = -w_{ji}^A$.

ANSWER: For a matrix with elements w_{ij} we have

$$w_{ij} = \frac{1}{2}[w_{ij} + w_{ji}] + \frac{1}{2}[w_{ij} - w_{ji}] = w_{ij}^S + w_{ij}^A$$

where $\mathbf{W}^S$ is symmetric

$$w_{ij}^S = \frac{1}{2}[w_{ij} + w_{ji}] = \frac{1}{2}[w_{ji} + w_{ij}] = w_{ji}^S$$

and $\mathbf{W}^A$ is anti-symmetric

$$w_{ij}^A = \frac{1}{2}[w_{ij} - w_{ji}] = -\frac{1}{2}[w_{ji} - w_{ij}] = -w_{ji}^A$$

2. Consider the 2^{nd} order terms in a 2^{nd} order polynomial in d dimensions, i.e. $\mathbf{x} = (x_1, \ldots, x_d)^T$.

$$\sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij} x_i x_j$$

Show that

$$\sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij} x_i x_j = \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}^S x_i x_j \tag{2}$$

i.e. there is no contribution from anti-symmetric matrix elements. This demonstrates that, without loss of generality, in problems involving (only) quadratic terms a matrix W can be taken to be *symmetric*, i.e. $W = W^S$.

ANSWER: We need to show that only the symmetric part contributes to the overall sum.

By using (1) we can split the sum in two parts: one for the symmetric matrix W^S and one for the anti-symmetric matrix W^A:

$$\begin{aligned}\sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}x_ix_j &= \sum_{i=1}^{d}\sum_{j=1}^{d}[w_{ij}^S + w_{ij}^A]x_ix_j \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}^S x_ix_j + \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}^A x_ix_j\end{aligned}$$

Using the anti-symmetric property of w^A in combination with the symmetry in $x_ix_j = x_jx_i$, the interchangeability of the summation sequence $\sum_{i=1}^{d}\sum_{j=1}^{d}\ldots = \sum_{j=1}^{d}\sum_{i=1}^{d}\ldots$, and the fact the that i and j are 'just' dummy-indices we find for the second sum S^A:

$$\begin{aligned}S^A &= \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}^A x_ix_j \\ &= \sum_{j=1}^{d}\sum_{i=1}^{d} w_{ji}^A x_jx_i \\ &= -\sum_{j=1}^{d}\sum_{i=1}^{d} w_{ij}^A x_jx_i \\ &= -\sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}^A x_ix_j \\ &= -S^A\end{aligned}$$

and $S^A = -S^A \Leftrightarrow S^A = 0$. Therefore the net contribution from the anti-symmetric sum is zero, and so

$$\sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}x_ix_j = \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}^S x_ix_j$$

Note: This property will feature prominently when the multivariate Gaussian distribution is discussed.

3. Show that the previous statement can also be stated in matrix notation as

$$\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{W}^S\mathbf{x} \tag{3}$$

with $\mathbf{W}^S = \frac{1}{2}\left(\mathbf{W} + \mathbf{W}^{\mathrm{T}}\right)$, the symmetric part of matrix $\mathbf{W}$.

ANSWER: From the definition of matrix multiplication, and the fact that a vector is just a

single column matrix, the left hand side of (3) is compactly written as

$$\begin{aligned}
\sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}x_i x_j &= \sum_{i=1}^{d}\sum_{j=1}^{d}[w_{ij}^S + w_{ij}^A]x_i x_j \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}^S x_i x_j + \underbrace{\sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}^A x_i x_j}_{=0} \\
&= \sum_{i=1}^{d} x_i \sum_{j=1}^{d} w_{ij}^S x_{j1} \\
&= \sum_{i=1}^{d} x_{1i}\left(\mathbf{W}^S\mathbf{x}\right)_{i1} \\
&= \left(\mathbf{x}^{\mathrm{T}}\mathbf{W}^S\mathbf{x}\right)_{11} \\
&= \mathbf{x}^{\mathrm{T}}\mathbf{W}^S\mathbf{x}
\end{aligned}$$

where the last step follows from the fact that the overall result is a scalar, i.e. a single cell matrix. The right hand side goes similar, after substituting the definition for the symmetric part of a matrix.

Exercise 2

In exercise 8, week 1, we considered the regression problem of approximating a data set of N input/output pairs $\{x_n, t_n\}$ by a polynomial function of the form

$$y(x;\mathbf{w}) = w_0 + w_1 x + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{4}$$

Applying the familiar steps (define an error measure, calculate gradient, set equal to zero and solve the equations), it could be shown that the squared error loss $E(\mathbf{w})$ for an M-th order polynomial was minimal when the weight coefficients $\mathbf{w}$ satisfied the following set of coupled equations

$$\sum_{j=0}^{M} w_j A_{ij} = T_i \tag{5}$$

with A_{ij} and T_i defined as

$$A_{ij} = \sum_{n=1}^{N} x_n^{i+j} \qquad T_i = \sum_{n=1}^{N} t_n x_n^i. \tag{6}$$

1. Verify that for a single data point $\{x_1, t_1\}$ the optimal solution for a first order polynomial through the origin takes the form

$$w_1 = \frac{1}{A_{11}} T_1 \tag{7}$$

ANSWER: A first order polynomial through the origin implies an equation of the form $y(x;\mathbf{w}) = w_1 x$, i.e. $w_0 = 0$. That means that, with $M = 1$, (5) reduces to

$$\sum_{j=1}^{1} w_j A_{ij} = w_1 A_{i1} = T_i \tag{8}$$

This holds for both equations, i.e. $i = 0$ and $i = 1$, and so choosing the latter and dividing both sides by A_{11} gives the result to show.

2. Show that for an arbitrary data set $\{x_n, t_n\}$ the optimal solution for an M-th order polynomial takes the form

$$\mathbf{w} = \mathbf{A}^{-1}\mathbf{T} \tag{9}$$

ANSWER:

Rewriting summation as matrix multiplication the set of equations (5) becomes

$$\begin{aligned} \sum_{j=0}^{M} w_j A_{ij} &= \sum_{j=0}^{M} A_{ij} w_{j1} \\ &= \big(\mathbf{Aw}\big)_{i1} = T_{i1} \end{aligned}$$

Combining all i components into matrix form this corresponds to

$$\mathbf{Aw} = \mathbf{T}$$

As left-multiplying by $\mathbf{A}^{-1}$ gives $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, the identity matrix, we get back equation (9):

$$\mathbf{w} = \mathbf{A}^{-1}\mathbf{T}$$

3. One technique that is often used to control the over-fitting phenomenon is *regularization*. Consider adding a penalty term to the squared error loss that takes the form of the sum-of-squares of all coefficients. The error function becomes:

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n;\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}||\mathbf{w}||^2, \tag{10}$$

where $||\mathbf{w}||^2 = \mathbf{w}^{\mathrm{T}}\mathbf{w} = \sum_{j=0}^{M} w_j^2$. Write down the set of coupled linear equations for the modified error function, analogous to the case without regularization:

$$\sum_{j=0}^{M} w_j \tilde{A}_{ij} = \tilde{T}_i \tag{11}$$

Build on the results from exercise 8, week 1. Compare $\tilde{A}_{ij}$ and $\tilde{T}_i$ to A_{ij} and T_i.

ANSWER:

In exercise 8, week 1, we determined the following:

$$\begin{aligned} \frac{\partial E}{\partial w_i} &= \sum_{n=1}^{N}\Big(\sum_{j=0}^{M} w_j x_n^{i+j} - t_n x_n^i\Big) \\ &= \sum_{n=1}^{N}\sum_{j=0}^{M} w_j x_n^{i+j} - \sum_{n=1}^{N} t_n x_n^i \\ &= \sum_{j=0}^{M}\sum_{n=1}^{N} x_n^{i+j} w_j - \sum_{n=1}^{N} t_n x_n^i \\ &= \sum_{j=0}^{M} A_{ij} w_j - T_i \end{aligned}$$

$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{\lambda}{2}||\mathbf{w}||^2 \implies$

$$\begin{aligned}
\frac{\partial \tilde{E}}{\partial w_i} &= \frac{\partial E}{\partial w_i} + \lambda w_i \\
&= \sum_{j=0}^{M} A_{ij} w_j - T_i + \lambda w_i \\
&= \sum_{j=0}^{M} A_{ij} w_j - T_i + \lambda \sum_{j=0}^{M} \delta_{ij} w_j \\
&= \sum_{j=0}^{M} (A_{ij} + \lambda \delta_{ij}) w_j - T_i \\
&= \sum_{j=0}^{M} w_j \tilde{A}_{ij} - \tilde{T}_i
\end{aligned}$$

We conclude that $\tilde{A}_{ij} = A_{ij} + \lambda \delta_{ij}$ and $\tilde{T}_i = T_i$.

Exercise 3

(see Bishop, eq.C.8 and C.9) The trace $\mathsf{Tr}\,(\mathbf{A})$ of a square matrix $\mathbf{A}$ is defined as the sum of the elements on the main diagonal:

$$\mathsf{Tr}\,(\mathbf{A}) = \sum_{i=1}^{N} A_{ii} \tag{12}$$

1. Prove by writing out in terms of indices that

$$\mathsf{Tr}\,(\mathbf{AB}) = \mathsf{Tr}\,(\mathbf{BA}) \tag{13}$$

ANSWER: $\mathsf{Tr}\,(\mathbf{A}) = \sum_i A_{ii}$, so

$$\mathsf{Tr}\,(\mathbf{AB}) = \sum_i \sum_k A_{ik} B_{ki} = \sum_k \sum_i B_{ki} A_{ik} = \mathsf{Tr}\,(\mathbf{BA})$$

2. Show that from this symmetry it follows that the trace is *cyclic*:

$$\mathsf{Tr}\,(\mathbf{ABC}) = \mathsf{Tr}\,(\mathbf{CAB}) = \mathsf{Tr}\,(\mathbf{BCA}) \tag{14}$$

ANSWER: Use

$$\mathbf{ABC} = \mathbf{A}(\mathbf{BC})$$

and take $(\mathbf{BC})$ to be a single matrix in the trace. The symmetry property (13) then implies:

$$\mathsf{Tr}\,(\mathbf{A}(\mathbf{BC})) = \mathsf{Tr}\,((\mathbf{BC})\mathbf{A})$$

etc.

Exercise 4

(see Bishop, eq.C.20) The derivative of a matrix $\mathbf{A}$ with elements A_{ij} depending on x is the matrix $\partial\mathbf{A}/\partial x$ with elements $\partial A_{ij}/\partial x$. Show, by writing out in elements, that

$$\frac{\partial}{\partial x}(\mathbf{AB}) = \frac{\partial \mathbf{A}}{\partial x}\mathbf{B} + \mathbf{A}\frac{\partial \mathbf{B}}{\partial x} \tag{15}$$

ANSWER:

$$\frac{\partial}{\partial x}(\sum_k A_{ik}B_{kj}) = \sum_k \frac{\partial A_{ik}}{\partial x}B_{kj} + \sum_k A_{ik}\frac{\partial B_{kj}}{\partial x}$$

Exercise 5

By repeatedly applying the product rule, show that

$$p(X, Y, Z) = p(Z|Y, X)p(Y|X)p(X) \tag{16}$$

ANSWER: The famous **product rule of probability**:

$$p(X, Y) = p(X|Y)p(Y)$$

(or: joint = conditional times marginal), so

$$p(X, Y, Z) = p(Z|Y, X)p(Y, X) = p(Z|Y, X)p(Y|X)p(X)$$

Exercise 6

Assume $p(Y) > 0$. Two equivalent criteria for independence are:

$$p(X, Y) = p(X)p(Y) \tag{17}$$
$$p(X|Y) = p(X) \tag{18}$$

Show that (17) implies (18) and vice versa. (When does the assumption $p(Y) > 0$ come into play?)

ANSWER: Use the fact that, by definition, the product rule of probability always holds. From (17) and the product rule it follows that $p(X|Y)p(Y) = p(X)p(Y)$. If also $p(Y) > 0$ then both sides can be divided by $p(Y)$ to give (18). (Here the assumption is crucial, since $a \cdot 0 = b \cdot 0$ does not imply that $a = b$.)

If (18) holds, then both sides can be multiplied by $p(Y)$; the product rule then immediately implies (17). (This is true even if $p(Y) = 0$).

Exercise 7

Suppose we have a box containing 8 apples and 4 grapefruit, and another box that contains 15 apples and 3 grapefruit. One of the boxes is selected at random ('50-50'), and then a piece of fruit is picked from the chosen box, again with equal probability for each item in the box.

1. Calculate the probability of selecting an apple.

   ANSWER: The **sum rule of probability** is defined as

   $$p(X) = \sum_Y p(X, Y)$$

   In combination with the product rule, with F representing the type of fruit (apple or grapefruit) and B the selected box (1 or 2), this can be written in the form

   $$p(F) = \sum_B p(F|B)p(B)$$

   Applying this to the case at hand we find

   $$\begin{aligned} p(F = a) &= p(B = 1)p(F = a|B = 1) + p(B = 2)p(F = a|B = 2) \\ &= \frac{1}{2}\left(\frac{8}{12} + \frac{15}{18}\right) = 0.75 \end{aligned}$$

2. The piece of fruit turns out to be an apple indeed. Use Bayes' (or Bayes's) rule to calculate the probability that it came from the first box.

   ANSWER: The posterior probability that box 1 was chosen, i.e. after observing the apple, is

   $$p(B = 1|F = a) = \frac{p(B = 1)p(F = a|B = 1)}{p(F = a)} = \frac{\frac{1}{2}\frac{8}{12}}{\frac{3}{4}} = \frac{4}{9} = 0.444...$$

3. The apple is replaced, and from the *same* box another piece of fruit is selected at random. What is the probability that this second pick is also an apple? (Note: same box, but *not* necessarily the first.)

   ANSWER:
   We found that the posterior probability on box 1, given that the first pick was an apple, is $\frac{4}{9}$ and so for box 2 it must be the remaining $\frac{5}{9}$. The probability that the next piece of fruit is again an apple is therefore

   $$\begin{aligned} p(F_2 = a|F_1 = a) &= p(F_2 = a|B = 1)p(B = 1|F_1 = a) + p(F_2 = a|B = 2)p(B = 2|F_1 = a) \\ &= \frac{4}{9}\frac{8}{12} + \frac{5}{9}\frac{15}{18} = 0.7593 \end{aligned}$$

   where we use that $p(F_2|B) = p(F_1|B) = p(F|B)$, since both pieces were selected *at random* from the same box.

Exercise 8 – MatLab basics

Before making this assignment, it is strongly suggested to work through some MatLab tutorials, for example `http://www.math.utah.edu/lab/ms/matlab/matlab.html` and `http://www.cyclismo.org/tutorial/matlab/` to learn or recap the basic MatLab syntax.

In MatLab or in GNU Octave, assign the following variables:

$$
\begin{aligned}
c &= 5 \\
x &= (1,2,3)^T \qquad \text{(a column vector)} \\
y &= (3,4,5,6) \qquad \text{(a row vector)} \\
z &= (4,5,6)^T \qquad \text{(a column vector)} \\
A &= \begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 3 \\ 4 & 6 & 1 \\ 0 & 0 & 1 \end{pmatrix} \\
B &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}
\end{aligned}
$$

and let MatLab calculate:

1. $\sum_{i=1}^{3} x_i$ (once using a for loop, and once using the sum command)
2. $\sum_{j=1}^{3} x_j$ (once using a for loop, and once using the sum command)
3. $\prod_{i=1}^{3} x_i$ (once using a for loop, and once using the prod command)
4. $\prod_{i=1}^{3} x_i$ (using the sum, exp and log commands)
5. $\sum_{j=1}^{3}(5x_j)$
6. $\sum_{j=1}^{3}(cx_j)$
7. $c\sum_{j=1}^{3} x_j$
8. $\sum_{i=1}^{3}(x_i + z_i)$
9. $\sum_{i=1}^{3} x_i + \sum_{i=1}^{3} z_i$
10. $\|x\| = \sqrt{\sum_{i=1}^{3} x_i^2}$
11. Ax
12. $\sum_{j=1}^{3} A_{ij}x_j$ for $i = 1, \ldots, 4$ (using nested for loops)
13. yA
14. yAx
15. $x^T A^T y^T$
16. AB
17. $B^T A^T$
18. $(AB)^T$
19. $\sum_{j=1}^{3} A_{ij}B_{jk}$ for $i = 1, \ldots, 4$ and $k = 1, 2$ (using nested for loops)
20. $\sum_{i=1}^{4}\sum_{j=1}^{3} A_{ij}$ (once using nested for loops, once using the sum command)
21. $\sum_{j=1}^{3}\sum_{i=1}^{4} A_{ij}$ (once using nested for loops, once using the sum command)

22. Write a recursive MatLab function to calculate $n!$ (remember that $n! = 1 \cdot 2 \cdot \cdots \cdot n$).

23. Write a single MatLab expression to calculate $n!$ (hint: you can use prod).

24. Write a lambda expression for the function $x \mapsto \sin(cx^2)$ and use it to calculate $\sum_{i=1}^{4} \sin(cy_i^2)$.

ANSWER:

```
c = 5
x = [1;2;3]
y = [3,4,5,6]
z = [4;5;6]
A = [1 2 0; 0 2 3; 4 6 1; 0 0 1]
B = [1 0; 0 1; 1 1]

% 1.
S = 0; for i=1:3; S = S + x(i); end; S
S = sum(x)

% 2.
S = 0; for j=1:3; S = S + x(j); end; S
S = sum(x)

% 3.
S = 1; for i=1:3; S = S * x(i); end; S
S = prod(x)

% 4.
S = exp(sum(log(x)))

% 5.
sum(5 * x)

% 6.
sum(c * x)

% 7.
c * sum(x)

% 8.
sum(x + z)

% 9.
sum(x) + sum(z)

% 10.
sqrt(sum(x.^2))

% 11.
A * x

% 12
S = zeros(4,1);
for i=1:4
  for j=1:3
    S(i) = S(i) + A(i,j) * x(j);
  end
end
```

```
S

% 13.
y * A

% 14.
y * A * x

% 15.
x' * A' * y'

% 16.
A * B

% 17.
B' * A'

% 18.
(A * B)'

% 19.
AB = zeros(4,2);
for i=1:4
  for k=1:2
    for j=1:3
      AB(i,k) = AB(i,k) + A(i,j) * B(j,k);
    end
  end
end
AB

% 20.
S = 0;
for i=1:4
  for j=1:3
    S = S + A(i,j);
  end
end
S
sum(sum(A))

% 21.
S = 0;
for j=1:3
  for i=1:4
    S = S + A(i,j);
  end
end
S
sum(sum(A))

% 22. in a file called fac.m:
function nfac = fac(n)
  if n < 0
    error 'n should be positive';
  elseif n == 0
    nfac = 1;
  else
```

```
    nfac = n * fac(n-1);
  end
return

% 23.
fac = @(n) prod([1:n]);

% 24.
f = @(x) sin(c * x.^2);
sum(f(y))
```

Exercise 9

Consider a probability density $p_x(x)$ defined over a continuous variable x, and suppose that we make a nonlinear change of variable using $x = g(y)$, so that the density transforms according to

$$p_y(y) = p_x(x)\left|\frac{dx}{dy}\right| = p_x\big(g(y)\big)\big|g'(y)\big|. \tag{19}$$

Assume that this nonlinear change of variables is monotonically increasing, i.e., $g'(y) > 0$ for all y. By differentiating this relationship, show that the location $\hat{y}$ of the maximum of the density in y is not in general related to the location $\hat{x}$ of the maximum of the density over x by the simple functional relation $\hat{x} = g(\hat{y})$ as a consequence of the Jacobian factor. This shows that the maximum of a probability density (in contrast to the simple function) is dependent on the choice of variable. Verify that, in the case of a linear transformation, the location of the maximum transforms the same way as the variable itself.

ANSWER: We are often interested in finding the most probable value for some quantity. In the case of probability distributions over discrete variables this poses little problem. However, for continuous variables there is a subtlety arising from the nature of probability densities and the way they transform under non-linear changes of variable.

Consider first the way a function $f(x)$ behaves when we change to a new variable y where the two variables are related by $x = g(y)$. This defines a new function of y given by

$$\tilde{f}(y) = f\big(g(y)\big). \tag{20}$$

Suppose $f(x)$ has a mode (i.e. a maximum) at $\hat{x}$ so that $f'(\hat{x}) = 0$. The corresponding mode of $\tilde{f}(y)$ will occur for a value $\hat{y}$ obtained by differentiating both sides of (20) with respect to y

$$\tilde{f}'(\hat{y}) = f'\big(g(\hat{y})\big)g'(\hat{y}) = 0.$$

As $g'(\hat{y}) \neq 0$, we have $f'\big(g(\hat{y})\big) = 0$. However, we know that $f'(\hat{x}) = 0$, and so we see that the locations of the mode expressed in terms of each of the variables x and y are related by $\hat{x} = g(\hat{y})$, as one would expect. Thus, finding a mode with respect to the variable x is completely equivalent to first transforming to the variable y, then finding a mode with respect to y, and then transforming back to x.

Now consider the behaviour of a probability density $p_x(x)$ under the change of variables $x = g(y)$, where the density with respect to the new variable is $p_y(y)$ and is given by (19). As $g' > 0$, (19) can be written

$$p_y(y) = p_x\big(g(y)\big)g'(y).$$

Differentiating both sides with respect to y then gives

$$p'_y(y) = p'_x\big(g(y)\big)\big(g'(y)\big)^2 + p_x\big(g(y)\big)g''(y). \tag{21}$$

Due to the presence of the second term on the right hand side of (21) the relationship $\hat{x} = g(\hat{y})$ no longer holds. Thus the value of x obtained by maximizing $p_x(x)$ will not be the value obtained by transforming to $p_y(y)$ then maximizing with respect to y and then transforming back to x. This causes modes of densities to be dependent on the choice of variables. In the case of linear transformation, the second term on the right hand side of (21) vanishes, and so the location of the maximum transforms according to $\hat{x} = g(\hat{y})$.