Statistical Machine Learning 2016

Exercises, 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 $w_{ij}^A = -w_{ji}^A$.

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. $W = W^S$.

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 $\mathbf{W}$.

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

2. Show that for an arbitrary data set $\{x_n, t_n\}$ the optimal solution for an M-th order polynomial takes the form
$$\mathbf{w} = \mathbf{A}^{-1}\mathbf{T} \tag{8}$$

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:

$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n; \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} ||\mathbf{w}||^2, \tag{9}$$

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:

$$\sum_{j=0}^{M} w_j \tilde{A}_{ij} = \tilde{T}_i \tag{10}$$

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

$$\mathsf{Tr}(\mathbf{A}) = \sum_{i=1}^{N} A_{ii} \tag{11}$$

1. Prove by writing out in terms of indices that

$$\mathsf{Tr}(\mathbf{AB}) = \mathsf{Tr}(\mathbf{BA}) \tag{12}$$

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

$$\mathsf{Tr}(\mathbf{ABC}) = \mathsf{Tr}(\mathbf{CAB}) = \mathsf{Tr}(\mathbf{BCA}) \tag{13}$$

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

$$\frac{\partial}{\partial x}(\mathbf{AB}) = \frac{\partial \mathbf{A}}{\partial x}\mathbf{B} + \mathbf{A}\frac{\partial \mathbf{B}}{\partial x} \tag{14}$$

Exercise 5

By repeatedly applying the product rule, show that

$$p(X, Y, Z) = p(Z|Y, X)p(Y|X)p(X) \tag{15}$$

Exercise 6

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

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

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

Show that (16) implies (17) and vice versa. (When does the assumption $p(Y) > 0$ come into play?)

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.
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.
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.)

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 $\sum_{i=1}^{4} \sin(cy_i^2)$.

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

$$p_y(y) = p_x(x) \left| \frac{dx}{dy} \right| = p_x\big(g(y)\big)\big|g'(y)\big|. \tag{18}$$

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.