Statistical Machine Learning 2016

Exercises and answers, week 1

Exercise 1

In analyzing problems in which a sigma-summation symbol is involved, it is sometimes helpful to write out the sum. By writing out the sum, I mean e.g.,

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

or more general

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n \ .$$

- Show, by explicitly writing out the sums, rearranging terms, and using brackets where needed, that the following four equations hold:

$$\sum_{i=1}^{3} (a x_i) = a\Big(\sum_{i=1}^{3} x_i\Big) \tag{1}$$

$$\sum_{i=1}^{3} \Big(\sum_{j=1}^{2} a_{ij}\Big) = \sum_{j=1}^{2} \Big(\sum_{i=1}^{3} a_{ij}\Big) \tag{2}$$

$$\sum_{i=1}^{3} \Big(\sum_{j=1}^{2} x_i y_j\Big) = \Big(\sum_{i=1}^{3} x_i\Big)\Big(\sum_{j=1}^{2} y_j\Big) \tag{3}$$

$$\sum_{i=1}^{3} a = 3a \tag{4}$$

ANSWER:
Show (1):

$$\begin{aligned}
\sum_{i=1}^{3} \big(a x_i\big) &= a x_1 + a x_2 + a x_3 \\
&= a(x_1 + x_2 + x_3) \\
&= a\Big(\sum_{i=1}^{3} x_i\Big)
\end{aligned}$$

Show (2):

$$\begin{aligned}\sum_{i=1}^{3}\Big(\sum_{j=1}^{2}a_{ij}\Big) &= \sum_{i=1}^{3}(a_{i1}+a_{i2})\\ &= (a_{11}+a_{12})+(a_{21}+a_{22})+(a_{31}+a_{32})\\ &= (a_{11}+a_{21}+a_{31})+(a_{12}+a_{22}+a_{32})\\ &= \sum_{j=1}^{2}(a_{1j}+a_{2j}+a_{3j})\\ &= \sum_{j=1}^{2}\Big(\sum_{i=1}^{3}a_{ij}\Big)\end{aligned}$$

Show (3):

$$\begin{aligned}\sum_{i=1}^{3}\Big(\sum_{j=1}^{2}x_iy_j\Big) &= (x_1y_1+x_1y_2)+(x_2y_1+x_2y_2)+(x_3y_1+x_3y_2)\\ &= x_1(y_1+y_2)+x_2(y_1+y_2)+x_3(y_1+y_2)\\ &= (x_1+x_2+x_3)(y_1+y_2)\\ &= \Big(\sum_{i=1}^{3}x_i\Big)\Big(\sum_{j=1}^{2}y_j\Big)\end{aligned}$$

Show (4):

$$\begin{aligned}\sum_{i=1}^{3}a &= a+a+a\\ &= 3a\end{aligned}$$

Exercise 2

Calculate the gradient ∇f of the following functions $f(\mathbf{x})$. In the left column, $\mathbf{x}=(x_1,x_2,x_3)$. In the right column, $\mathbf{x}=(x_1,\ldots,x_n)$.

a) $f(x_1,x_2,x_3)=a_1x_1+a_2x_2+a_3x_3$	e) $f(\mathbf{x})=\sum_{i=1}^{n}a_ix_i$
b) $f(x_1,x_2,x_3)=x_2$	f) $f(\mathbf{x})=x_i$
c) $f(x_1,x_2,x_3)=x_1x_2x_3$	g) $f(\mathbf{x})=\prod_{i=1}^{n}x_i$
d) $f(x_1,x_2,x_3)=x_1^{k_1}x_2^{k_2}x_3^{k_3}$	h) $f(\mathbf{x})=\prod_{i=1}^{n}x_i^{k_i}$

Note: often it suffices to write down the partial derivative $\partial f/\partial x_j$ (Can you tell why?).

ANSWER: a) (a_1,a_2,a_3), in other words $\partial f/\partial x_i=a_i,\ i=1\ldots3$
b) $(0,1,0)$, in other words $\partial f/\partial x_j=\delta_{2j},\ j=1\ldots3$ (Kronecker delta, see slides)
c) (x_2x_3,x_1x_3,x_1x_2)
d) $(k_1x_1^{k_1-1}x_2^{k_2}x_3^{k_3},k_2x_1^{k_1}x_2^{k_2-1}x_3^{k_3},k_3x_1^{k_1}x_2^{k_2}x_3^{k_3-1})$ (where the $k_ix_i^{k_i-1}$ is understood as 0 if $k_i=0$)
e) $(a_1,\ldots,a_n)$ in other words $\partial f/\partial x_j=a_j$

f) $(\delta_{i1}, \ldots, \delta_{in})$ in other words $\partial f/\partial x_j = \delta_{ij}$
g) Note that $\prod_{i=1}^n x_i = (\prod_{i=1, i\neq j}^n x_i) x_j$, so $\partial f/\partial x_j = \prod_{i=1, i\neq j}^n x_i$
h) $\partial f/\partial x_j = k_j x^{k_j - 1} \prod_{i=1, i\neq j}^n x_i^{k_i}$
To describe a vector say $\vec{u} = (u_1, u_2, \ldots, u_j, \ldots, u_n)$, it suffices to give the expresssion of an arbitrary component u_j. So u_j is some expression that contains j's. All components and so the complete vector can then be reconstructed by filling in the appropriate component number for j. E.g. if you look for u_2, take the general expression for u_j and substitute all the j's by a 2. Now the gradient ∇f is also a vector. Its j-th component is just $\partial f/\partial x_j$, which is therefore sufficient to describe the vector ∇f.

In many cases, it is convenient to only write down the abstract j component. However, it should be remembered that the gradient is an object with n components, and that it is sometimes more convenient to write down all the components. I think this could be argued for e.g. the gradient in b), $\nabla f = (0, 1, 0)$.

Exercise 3

The function

$$f(x, y) = 2x^2 - xy + y^2 - x + y + 5.5 \tag{5}$$

has a unique minimum (x^*, y^*). Calculate this point.

ANSWER: Partial derivatives of f are given by

$$\begin{aligned}\frac{\partial f}{\partial x} &= 4x - y - 1 \\ \frac{\partial f}{\partial y} &= 2y - x + 1\end{aligned}$$

Setting equal to zero yields two equations for x and y. Solve the first to get: $y = 4x - 1$. Substituting in the second then gives: $8x - 2 - x + 1 = 7x - 1 = 0 \Rightarrow x^* = 1/7$, and so $y^* = -3/7$.

(As a side remark: it is indeed a *minimum* since the Hessian, the matrix of second order partial derivatives, is positive definite, meaning that $x^T M x > 0$ for all vectors x. An equivalent statement is that the eigenvalues λ_i of the matrix M are all positive.)

Exercise 4

Calculate the minimum x^* of the following two functions.

$$f(x) = \sum_{i=1}^{n} (x - a_i)^2 \tag{6}$$

ANSWER:

$$\begin{aligned}\frac{df(x)}{dx} &= 2\sum_{i=1}^{n}(x - a_i) = 0 \\ &\Rightarrow \sum_{i=1}^{n} x = \sum_{i=1}^{n} a_i \\ &\Rightarrow nx = \sum_{i=1}^{n} a_i \\ &\Rightarrow x = \frac{1}{n}\sum_{i=1}^{n} a_i\end{aligned}$$

so x is the mean of the a_i's.

There are several things to note:
1) derivative of a sum is a sum of derivatives
2) x has no subindex i. Therefore

$$\sum_{i=1}^{n} x = \underbrace{x + x + \ldots + x}_{n \text{ times}} = nx$$

As a side remark: this minimization can be seen as a least square problem: given data a_i, which x gives the best fit such that the sum of the squares of the errors $(x - a_i)$ is minimal. The solution is the data mean.

$$f(x) = \sum_{i=1}^{n} \alpha_i (x - a_i)^2 \quad (\text{with } \alpha_i > 0) \tag{7}$$

ANSWER:

$$\begin{aligned} \frac{df(x)}{dx} &= 2\sum_{i=1}^{n} \alpha_i (x - a_i) = 0 \\ &\Rightarrow \sum_{i=1}^{n} \alpha_i x = \sum_{i=1}^{n} \alpha_i a_i \\ &\Rightarrow x = \frac{\sum_{i=1}^{n} \alpha_i a_i}{\sum_{i=1}^{n} \alpha_i} \end{aligned}$$

Side remark: This minimization can be seen as a weighted least square problem: given data a_i, which x gives the best fit such that the weighted sum of the squares of the errors $(x - a_i)$ is minimal. The solution is the weighted average so here x is the weighted average of a_i with weights α_i. The factor in the denominator (noemer) is for normalization (just as the n is in the previous case).

Exercise 5

Calculate the gradient ∇f of

$$f(\vec{h}) = \sum_{i=1}^{n} p_i h_i - \ln\Big(\sum_{i=1}^{n} \exp(h_i)\Big) \tag{8}$$

ANSWER:
$\vec{h}$ is a vector of n components $(h_1, \ldots, h_n)$. The function $f(\vec{h})$ is a scalar function of these n components; the p_i are constants. The gradient ∇f is then the vector of partial derivatives of f w.r.t. each component h_j. Since

$$\frac{\partial}{\partial h_j}\left[\sum_{i=1}^{n} p_i h_i\right] = p_j$$

and

$$\frac{\partial}{\partial h_j}\left[\sum_{i=1}^{n} \exp(h_i)\right] = \exp(h_j)$$

application of the chain rule to (8) gives

$$\frac{\partial f}{\partial h_j} = p_j - \frac{\exp(h_j)}{\sum_{i=1}^{n} \exp(h_i)}$$

Side remark: this f is related to a so-called likelihood function (will be treated later in the course).

Exercise 6

Compute the minimum x^* of

$$f(x) = a \ln(x) + \frac{b}{2x^2} \tag{9}$$

with $a > 0$, $b > 0$ en $x > 0$. Express your answer in terms of a en b. (Note: $\ln(x)' = 1/x$).

ANSWER: Calculate gradient (slope) of f, set equal to zero and solve for x^*

$$\begin{aligned} \frac{a}{x} - bx^{-3} = 0 \quad &\Rightarrow \quad a - bx^{-2} = 0 \\ &\Rightarrow \quad ax^2 - b = 0 \\ &\Rightarrow \quad x = \sqrt{b/a} \end{aligned}$$

Side remark: this f is also related to (another) likelihood function (will also be treated later in the course).

Exercise 7

(see Bishop, appendix C, eq.C.1) An $N \times M$ matrix $\mathbf{A}$ has elements A_{ij} (with i the row- and j the columnindex). The transposed matrix $\mathbf{A}^T$ has elements $(\mathbf{A}^T)_{ij} = A_{ji}$. By writing out the matrix product using index notation show that

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T \tag{10}$$

Hint: $\mathbf{C} = \mathbf{AB}$ corresponds to $C_{ij} = \sum_{k=1}^{M} A_{ik} B_{kj}$

ANSWER: $(\mathbf{A})_{ij} = A_{ij}$, $(\mathbf{A}^T)_{ij} = A_{ji}$, $(\mathbf{AB})_{ij} = \sum_k A_{ik} B_{kj}$ so

$$\begin{aligned} ((\mathbf{AB})^T)_{ij} &= (\mathbf{AB})_{ji} = \sum_k A_{jk} B_{ki} = \sum_k B_{ki} A_{jk} \\ &= \sum_k (\mathbf{B}^T)_{ik} (\mathbf{A}^T)_{kj} = (\mathbf{B}^T \mathbf{A}^T)_{ij} \end{aligned}$$

Exercise 8

(Exercise 1.1 from the Bishop book.) Consider the M-th order polynomial

$$y(x; \mathbf{w}) = w_0 + w_1 x + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{11}$$

and the error function

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n;\mathbf{w}) - t_n\}^2 \tag{12}$$

with x_n, t_n the input/output pairs from the data set. Define the error per data point as

$$E_n(\mathbf{w}) = \frac{1}{2}\{y(x_n;\mathbf{w}) - t_n\}^2 \tag{13}$$

(so $E = \sum_{n=1}^{N} E_n$). Note that x = 1-dimensional, and that in this exercise the super-indices i, j represent 'power'.

1. Calculate the gradient of the error per data point E_n:

$$\nabla E_n \quad (= (\frac{\partial E_n}{\partial w_0}, \ldots, \frac{\partial E_n}{\partial w_M})^T\ ). \tag{14}$$

ANSWER: Use the chain rule on $E_n(\mathbf{w}) = \frac{1}{2}\{u(\mathbf{w})\}^2$ with $u(\mathbf{w}) = y(x_n;\mathbf{w}) - t_n$. Then for the components of the gradient

$$\begin{aligned}
\frac{\partial E_n}{\partial w_i} &= \frac{\partial E_n}{\partial u}\frac{\partial u}{\partial w_i} \\
&= u(\mathbf{w})\,\frac{\partial u(\mathbf{w})}{\partial w_i} \\
&= (y(x_n;\mathbf{w}) - t_n)\,\frac{\partial y(x_n;\mathbf{w}) - t_n}{\partial w_i} \\
&= \left(\sum_{j=0}^{M} w_j (x_n)^j - t_n\right)\frac{\partial}{\partial w_i}\left[\sum_{k=0}^{M} w_k x_n^k - t_n\right] \\
&= \left(\sum_{j=0}^{M} w_j x_n^j - t_n\right) x_n^i \\
&= \sum_{j=0}^{M} w_j x_n^{i+j} - t_n x_n^i
\end{aligned}$$

Note that x_n^i means: x_n to-the-power-of i. Note that in general $x^a x^b = x^{a+b}$, e.g. $2^3 2^4 = 2^7$

If you got this answer by direct differentiation e.g. by writing out the y's in terms of w's, without the use of an u, that is of course also ok.

2. Calculate the gradient of the total error E.

ANSWER: The total error E is the sum of the errors per datapoint E_n. Since the gradient is a linear function of its operands: $\nabla(f + g) = \nabla f + \nabla g$, the gradient of the total error is the sum of the gradients of the error per datapoint:

$$\nabla E = \sum_{n=1}^{N} \nabla E_n$$

with∇E_n as above. So,

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N}(\sum_{j=0}^{M} w_j x_n^{i+j} - t_n x_n^i)$$

3. Show that the partial derivatives can be written as

$$\frac{\partial E}{\partial w_i} = \sum_{j=0}^{M} A_{ij} w_j - T_i \tag{15}$$

with A_{ij} and T_i defined as

$$A_{ij} = \sum_{n=1}^{N} x_n^{i+j} \qquad T_i = \sum_{n=1}^{N} t_n x_n^i. \tag{16}$$

ANSWER: Substituting the result for the components of ∇E_n into (2) we have

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

4. When E is minimal it holds that $\nabla E = 0$ (i.e., all partial derivatives are zero). Using this, show that in the minimum of E the parameters $\mathbf{w}$ satisfy

$$\sum_{j=0}^{M} A_{ij} w_j = T_i. \tag{17}$$

ANSWER: In the last result, setting all partial derivatives equal to zero implies that when the error is minimal then

$$\sum_{j=0}^{M} A_{ij} w_j - T_i = 0 \;\Rightarrow\; \sum_{j=0}^{M} A_{ij} w_j = T_i.$$