Statistical Machine Learning 2018

Exercises, week 1

7 September 2018

TUTORIAL

Exercise 1

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) $f(x_1, x_2, x_3) = a_1 x_1 + a_2 x_2 + a_3 x_3$	e) $f(\mathbf{x}) = \sum_{i=1}^{n} a_i x_i$
b) $f(x_1, x_2, x_3) = x_2$	f) $f(\mathbf{x}) = x_i$
c) $f(x_1, x_2, x_3) = x_1 x_2 x_3$	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?).

Exercise 2

The function

$$f(x, y) = 2x^2 - xy + y^2 - x + y + 5.5 \tag{1}$$

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

Exercise 3

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

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

Exercise 4

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

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}}, \tag{2}$$

where $\mathbf{A}$ is a $M \times N$ matrix and $\mathbf{B}$ is a $N \times P$ matrix.

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

Exercise 5

(see Bishop, Exercise 1.1) Consider the M-th order polynomial

$$y(x; \mathbf{w}) = w_0 + w_1 x + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{3}$$

and the error function

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n; \mathbf{w}) - t_n\}^2 \tag{4}$$

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

$$E_n(\mathbf{w}) = \frac{1}{2} \{y(x_n; \mathbf{w}) - t_n\}^2 \tag{5}$$

(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:

$$\nabla E_n \quad (= (\frac{\partial E_n}{\partial w_0}, \ldots, \frac{\partial E_n}{\partial w_M})^T). \tag{6}$$

2. Calculate the gradient of the total error E.

3. Show that the partial derivatives can be written as

$$\frac{\partial E}{\partial w_i} = \sum_{j=0}^{M} A_{ij} w_j - T_i \tag{7}$$

with A_{ij} and T_i defined as

$$A_{ij} = \sum_{n=1}^{N} x_n^{i+j} \qquad T_i = \sum_{n=1}^{N} t_n x_n^i. \tag{8}$$

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

$$\sum_{j=0}^{M} A_{ij} w_j = T_i. \tag{9}$$

5. 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

$$w_1 = \frac{1}{A_{11}} T_1 \tag{10}$$

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

$$\mathbf{w} = \mathbf{A}^{-1} \mathbf{T} \tag{11}$$

7. 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:

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

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:

$$\sum_{j=0}^{M} w_j \tilde{A}_{ij} = \tilde{T}_i. \tag{13}$$

Compare $\tilde{A}_{ij}$ and $\tilde{T}_i$ to A_{ij} and T_i.

Exercise 6

In this exercise, we will have a closer look at the gradient descent algorithm for function minimization. When the function to be minimized is $E(\mathbf{x})$, the gradient descent iteration is

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \eta \nabla E(\mathbf{x}_n) \tag{14}$$

where $\eta > 0$ is the so-called learning-rate.

1. Consider the function $E(x) = \frac{\lambda}{2}(x-a)^2$ with parameters $\lambda > 0$, and a arbitrary.
   (a) Write down the gradient descent iteration rule. Verify that the minimum of E is a and that a is a fixed point[1] of the gradient descent iteration rule.
   (b) Show that the algorithm converges in one step if $\eta = 1/\lambda$.
   (c) Define $d_n = x_n - a$. Show that if $0 < \eta < 1/\lambda$, subsequent d_n's have the same signs. Also show that if $\eta > 1/\lambda$, subsequent d_n's have opposite signs.
   (d) The distance to the fixed point is $|d_n|$. Show that $|d_{n+1}| = |(1-\eta\lambda)||d_n|$. Show that this implies that the algorithm converges to the fixed point if $0 < \eta < 2/\lambda$, and that it diverges if $\eta > 2/\lambda$.
2. Consider now the function $E(x, y) = \frac{\lambda_1}{2}(x-a_1)^2 + \frac{\lambda_2}{2}(y-a_2)^2$ with parameters $0 < \lambda_1 < \lambda_2$, and a_i arbitrary.
   (a) Write down the gradient descent iteration rule. Verify that the minimum of E is a fixed point.
   (b) We want to find the learning rate η that leads to the fasted convergence in both x and y direction. This optimal learning rate is the one for which both $|1-\eta\lambda_1|$ and $|1-\eta\lambda_2|$ are as small as possible. For the optimal learning rate, the equation $|1-\eta\lambda_1| = |1-\eta\lambda_2|$ must therefore hold. Since $\lambda_1 < \lambda_2$, this can only hold if $\eta\lambda_1 < 1$ and $\eta\lambda_2 > 1$.
      - Show that solving the equation leads to $\eta^* = 2/(\lambda_2 + \lambda_1)$ (which is the optimal learning rate). What happens if η is smaller than the optimal value? What happens if it is larger?
   (c) What is the value of $|1-\eta^*\lambda_i|$ in both directions? What does this say about the applicability of gradient descent to functions with steep hills and flat valleys (i.e., if $\lambda_2 \gg \lambda_1$)?

[1] A fixed point x^* of an iteration $x_{n+1} = F(x_n)$ satisfies $x^* = F(x^*)$.

BONUS PRACTICE

Exercise 7

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:

$$\sum_{i=1}^{3} (a x_i) = a\big(\sum_{i=1}^{3} x_i\big) \tag{15}$$

$$\sum_{i=1}^{3} \big(\sum_{j=1}^{2} a_{ij}\big) = \sum_{j=1}^{2} \big(\sum_{i=1}^{3} a_{ij}\big) \tag{16}$$

$$\sum_{i=1}^{3} \big(\sum_{j=1}^{2} x_i y_j\big) = \big(\sum_{i=1}^{3} x_i\big)\big(\sum_{j=1}^{2} y_j\big) \tag{17}$$

$$\sum_{i=1}^{3} a = 3a \tag{18}$$

Exercise 8

Calculate the gradient ∇f of

$$f(\vec{h}) = \sum_{i=1}^{n} p_i h_i - \ln\Big(\sum_{i=1}^{n} \exp(h_i)\Big) \tag{19}$$

Exercise 9

Compute the minimum x^* of

$$f(x) = a \ln(x) + \frac{b}{2x^2} \tag{20}$$

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

Exercise 10

(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:

$$\mathsf{Tr}\,(\mathbf{A}) = \sum_{i=1}^{N} A_{ii} \tag{21}$$

1. Prove by writing out in terms of indices that

$$\mathsf{Tr}\left(\mathbf{AB}\right) = \mathsf{Tr}\left(\mathbf{BA}\right) \tag{22}$$

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

$$\mathsf{Tr}\left(\mathbf{ABC}\right) = \mathsf{Tr}\left(\mathbf{CAB}\right) = \mathsf{Tr}\left(\mathbf{BCA}\right) \tag{23}$$

Exercise 11

(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

$$\frac{\partial}{\partial x}(\mathbf{AB}) = \frac{\partial \mathbf{A}}{\partial x}\mathbf{B} + \mathbf{A}\frac{\partial \mathbf{B}}{\partial x} \tag{24}$$