Machine Learning 1 - Practice exercises 1

1 Basic Linear Algebra and Derivatives

Question 1.1

Let $\mathbf{A} = \begin{bmatrix} -7 & 8 & 1 \\ -4 & 3 & 5 \\ 7 & 7 & -8 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$

1. Compute $\mathbf{Ab}$.
2. Compute $\mathbf{b}^T\mathbf{A}$.
3. Compute the vector $\mathbf{c}$ for which $\mathbf{Ac} = \mathbf{b}$ through elimination.
4. Compute the inverse of $\mathbf{A}$, denoted by $\mathbf{A}^{-1}$.
5. Verify that in this case $\mathbf{A}^{-1}\mathbf{b} = \mathbf{c}$. Show that this must be the case for general $\mathbf{A}$, $\mathbf{b}$ and $\mathbf{c}$.

Solutions

1. $\begin{bmatrix} -7 + 2 \cdot 8 + 5 \\ -4 + 2 \cdot 3 + 5^2 \\ 7 + 2 \cdot 7 + -8 \cdot 5 \end{bmatrix} = \begin{bmatrix} 14 \\ 27 \\ -19 \end{bmatrix}$
2. $\big[-7 \cdot 1 + -4 \cdot 2 + 7 \cdot 5, 8 \cdot 1 + 3 \cdot 2 + 7 \cdot 5, 1 \cdot 1 + 5 \cdot 2 + 8 \cdot 5\big] = \big[20, 49, -29\big]$
3.

-7	8	1	1
-4	3	5	2
7	7	-8	5
0	15	-7	6
0	49	3	34
0	388	0	256

$$c_2 = \frac{256}{388} \approx 0.660$$
$$c_3 = \frac{34 - 49 \cdot \frac{256}{388}}{3} \approx 0.557$$
$$c_1 = \frac{2 - 5c_3 - 3c_2}{-4} \approx 0.691$$

$$c = \begin{bmatrix} 0.691 \\ 0.660 \\ 0.557 \end{bmatrix}$$

4.

-7	8	1	1	0	0
-4	3	5	0	1	0
7	7	-8	0	0	1
0	15	-7	1	0	1
0	49	3	0	7	4
0	388	0	3	49	31
0	0	388	-49	105	11
49	0	-59	0	-7	3
11	0	-37	3	-8	0
388	0	0	-59	71	37

$$\mathbf{A}^{-1} = \frac{1}{388} \begin{bmatrix} -59 & 71 & 37 \\ 3 & 49 & 31 \\ -49 & 105 & 11 \end{bmatrix}$$

Question 1.2

Find the derivative of the following functions with respect to x.

1. $\left(\frac{2}{x^2} + x^{-7} + x^3\right)^2$
2. $x^2\sqrt{e^{-\sqrt[3]{x}}}$
3. $x + \ln(x)$
4. $x\ln(\sqrt{x})$
5. $6(x^2 - 1)\sin(x)$
6. $\ln\left(\sqrt[3]{\frac{e^{3x}}{1 + e^{3x}}}\right)$

Partial derivatives:

7. What does it mean to build the partial derivative of a multivariate function?

Find the partial derivative of the following functions with respect to x,y, z

8. $f(x, y, z) = 2\ln(y - \exp(x^{-1}) - \sin(zx^2))$
9. $f(x, y, z) = \ln\left(\sqrt[\gamma]{z^\alpha y^\beta x^\gamma}\right)$

Solutions

1. $2(\frac{2}{x^2} + x^{-7} + x^3)(\frac{-4}{x^3} - 7x^{-8} + 3x^2) = \frac{2(x^5+1)^3(3x^5-7)}{x^{15}}$
2. $2x\sqrt{e^{-\sqrt[3]{x}}} - \frac{1}{6}\sqrt{e^{-\sqrt[3]{x}}}x^{4/3} = x\sqrt{e^{-\sqrt[3]{x}}}(2 - \frac{\sqrt[3]{x}}{6})$
3. $1/x + 1$
4. $\frac{1}{2}(1 + \ln(x))$
5. $6(-1 + x^2)\cos(x) + 12x\sin(x)$
6. $\frac{1}{\sqrt[3]{\frac{e^{3x}}{1+e^{3x}}}}\frac{1}{3}\sqrt[3/2]{\frac{e^{3x}}{1+e^{3x}}}\frac{1}{(e^{3x}+1)^2}3e^{3x} = \frac{1}{(1+e^{3x})}$
7. direction of steepest decent in one direction
8. $\frac{2(e^{(1/x)}/x^2 - 2xz\cos(x^2z))}{-\sin(x^2z) - e^{(1/x)} + y}, \frac{-2}{\sin(x^2z) + e^{(1/x)} - y}, \frac{2x^2\cos(x^2z)}{\sin(x^2z) + e^(1/x) - y}$
9. $\frac{1}{x}, \frac{\beta}{\gamma y}, \frac{\alpha}{z\gamma}$

Question 1.3

The following questions are good practice in manipulating vectors and matrices and they are very important for solving for posterior distributions.

Given the following expression:

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) + (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \mathbf{S}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)$$

where $\mathbf{x}$, $\boldsymbol{\mu}$, $\boldsymbol{\mu}_0$ are vectors and $\boldsymbol{\Sigma}^{-1}$ and $\mathbf{S}^{-1}$ are symmetric, *positive semi-definite* invertible matrices.

Answer the following questions:

1. Expand the expression and gather terms.
2. Collect all the terms that depend on $\boldsymbol{\mu}$ and those that do not.
3. Take the derivative with respect to $\boldsymbol{\mu}$, set to 0, and solve for $\boldsymbol{\mu}$.

Solutions

1. (1pt)

$$\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}^T\mathbf{S}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}_0^T\mathbf{S}^{-1}\boldsymbol{\mu}_0 - 2\boldsymbol{\mu}^T\mathbf{S}^{-1}\boldsymbol{\mu}_0$$

2. (1pt)

$$\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1} + \mathbf{S}^{-1}\right)\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{S}^{-1}\boldsymbol{\mu}_0\right) + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_0^T\mathbf{S}^{-1}\boldsymbol{\mu}_0$$

3. (1pt)

$$f(\boldsymbol{\mu}) = \boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1} + \mathbf{S}^{-1}\right)\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{S}^{-1}\boldsymbol{\mu}_0\right) + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_0^T\mathbf{S}^{-1}\boldsymbol{\mu}_0$$

We will use the convention that $\frac{\partial f(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}}$ is a row vector, since the chain rule works more naturally in this convention. (Note that the matrix cookbook uses the opposite convention and treats it as a column vector). You are free to use either notation, but you should be consistent and make sure you know how to apply the chain rule in the convention of your choice.

$$\frac{\partial f(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = 2\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1} + \mathbf{S}^{-1}\right) - 2\left(\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{S}^{-1}\boldsymbol{\mu}_0\right)^T = 0$$

Taking the transpose of the lefthand side and the righthand side, and using the fact that $\boldsymbol{\Sigma}^{-1}$ and $\mathbf{S}^{-1}$, are positive semi-definite, symmetric and invertible, we obtain

$$\begin{aligned}\left(\boldsymbol{\Sigma}^{-1} + \mathbf{S}^{-1}\right)\boldsymbol{\mu} &= \left(\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{S}^{-1}\boldsymbol{\mu}_0\right) \\ \boldsymbol{\mu} &= \left(\boldsymbol{\Sigma}^{-1} + \mathbf{S}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{S}^{-1}\boldsymbol{\mu}_0\right).\end{aligned}$$

2 Probability Theory

For these questions you will practice manipulating probabilities and probability density functions using the sum and product rules.

Question 2.1

Being a student in the Netherlands, you spend all your time in the cities of Amsterdam and Rotterdam. Based on your experience, the weather in Amsterdam is much nicer: the probability that it rains when you are in Amsterdam is 0.5, while the probability that it rains when in Rotterdam is 0.75. Amsterdam is where you spend most of your time: at any given moment, the probability that you are in Amsterdam is 0.8 and the probability that you are in Rotterdam is 0.2.

Based on the above:

1. Define the random variables and the values they can take on, both with symbols and numerically.
2. What is the probability that it does not rain when you are in Rotterdam?
3. What is the probability that it rains at your current location?
4. You wake up on the sidewalk, after a night out which you can't remember anything about but which clearly was not such a great idea. You can't recognize your surroundings, but you must be either in Amsterdam or Rotterdam. It is raining. What is the probability that you are in Amsterdam?

Solutions

1. W: weather {rainy,sunny}, L: location {Rotterdam,Amsterdam}
2. Let $W \in \{s, r\}$ be the weather (sunny or rainy) and $L \in \{r, a\}$ be the location (Rotterdam or Amsterdam). Then
   $P(W = s|L = r) = 1 - 0.75 = 0.25$.
3. $$\begin{aligned} P(W = r) &= P(L = r)P(W = r|L = r) + P(L = a)P(W = r|L = a) \\ &= 0.2 \times 0.75 + 0.8 \times 0.5 \\ &= 0.55 \end{aligned}$$
4. $$\begin{aligned} P(L = a|W = r) &= \frac{P(L = a)P(W = r|L = a)}{P(W = r)} \\ &= \frac{0.8 \times 0.5}{0.55} \\ &= 0.727 \end{aligned}$$

Question 2.2

According to a 1988 study, home pregnancy tests conducted by experienced technicians have a false negative rate of only 2.6%, while when carried out by consumers the rate goes up to 25%, due to improper use or misunderstanding of the instructions. Generally, false positive rates are rare events (0.01%) independent of the conductor. With this information, given a group of 10 000 women that could be evenly likely pregnant or not, 2000 are being assisted by a professional (Group A), and 8000 conduct the test on their own (Group B).

1. Define the random variables and the values they can take on.
2. How many women in each group do you expect being diagnosed as pregnant?
3. In each group, for how many do you expect a wrong result?

Solutions

1. (1pt) pregnant...$P \in \{true, false\}$, positively tested ... $T \in \{true, false\}$, group ... $G \in \{assisted, unassisted\}$
2. (1pt)

$$\begin{aligned} p(T = true|G = assisted) =& p(T = true|G = assisted, p = true)p(p = true)+ \\ & p(T = true|G = assisted, P = false)p(P = false) \\ =& (1 - 0.026)(0.5) + (0.0001)(0.5) = 0.48705 = 9741/20000 \end{aligned}$$

   We expect $2000 \cdot 0.48705 = 974,1 \approx 974$ women to be told that they are pregnant in the assisted group.

$$\begin{aligned} p(T = true|G = unassisted) =& p(T = true|G = unassisted, P = true)p(P = true)+ \\ & p(T = true|G = unassisted, p = false)p(p = false) \\ =& (1 - 0.25)(0.5) + (0.0001)(0.5) = 0.37505 = 7501/20000 \end{aligned}$$

   We expect $8000 \cdot 0.37505 = 3000.4 \approx 3000$ women to be told that they are pregnant in the unassisted group.

3. (1pt)

$$\begin{aligned} p =& p(T = true|G = assisted, P = false) * p(P = false)+ \\ & p(T = false|G = assisted, P = true) * p(P = true) \\ =& (0.026)(0.5) + (0.0001)(0.5) = 0,01305 = 261/20000 \end{aligned}$$

   We expect the test result to be wrong for 1.3% of the women = (26.1) in the assisted group.

$$\begin{aligned} p =& p(T = true|G = assisted, P = false) * p(P = false)+ \\ & p(T = false|G = assisted, P = true) * p(P = true) \\ =& (0.25)(0.5) + (0.0001)(0.5) = 0.12505 = 2501/20000 \end{aligned}$$

   We expect the test result to be wrong for 12.5% of the women = (1000.4) in the unassisted group.

Question 2.3

For this question you will compute the expression for the posterior parameter distribution for a simple data problem. Assume we observe N univariate data points $\{x_1, x_2, \ldots, x_N\}$. Further, we assume that they are generated by a Gaussian distribution with known variance σ^2, but unknown mean μ. Assume a prior Gaussian distribution over the unknown mean, i.e. $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$. When answering these questions, use $\mathcal{N}(a|b, c^2)$ to indicate a Gaussian (normal) distribution over a with mean b and variance c^2. You do not need to write down the explicit form of a guassian distribution.

1. Write down the general expression for a posterior distribution, using θ for the parameter, $\mathcal{D}$ for the data. Indicate the *prior*, *likelihood*, *evidence*, and *posterior*.

2. Write the posterior for this particular example. You do not need an analytic solution.

Solutions

1. (1pt)

$$\underbrace{p(\theta|\mathcal{D})}_{\text{posterior}} = \frac{p(\theta)p(\mathcal{D}|\theta)}{\int p(\theta)p(\mathcal{D}|\theta)d\theta}$$

$$= \frac{\overbrace{p(\theta)}^{\text{prior}}\overbrace{p(\mathcal{D}|\theta)}^{\text{likelihood}}}{\underbrace{p(\mathcal{D})}_{\text{evidence}}}$$

2. (1pt)

$$p(\theta|\mathcal{D}) = \frac{\mathcal{N}(\mu|\mu_0, \sigma_0^2)\prod_{n=1}^{N}\mathcal{N}(x_n|\mu, \sigma^2)}{\int \mathcal{N}(\mu|\mu_0, \sigma_0^2)\prod_{n=1}^{N}\mathcal{N}(x_n|\mu, \sigma^2)d\mu}$$