Exercise 1 - Eigenvectors and Eigenvalues

For a square matrix $\mathbf{A}$ of size $n \times n$, a vector $\mathbf{u}_i \neq 0$ which satisfies

$$\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{1}$$

is called a *eigenvector* of $\mathbf{A}$, and λ_i is the corresponding *eigenvalue*. For a matrix of size $n \times n$, there are n eigenvalues λ_i (which are not necessarily distinct).

Show that if $\mathbf{u}_1$ and $\mathbf{u}_2$ are eigenvectors with equal corresponding eigenvalues $\lambda_1 = \lambda_2$, then $\mathbf{u} = \alpha\mathbf{u}_1 + \beta\mathbf{u}_2$ is also an eigenvector with the same eigenvalue.

Solution

Because $\lambda_1 = \lambda_2$, we will write λ for simplicity. The result is obtained by applying the definition of eigenvalues and distributivity via

$$\mathbf{A}\mathbf{u} = \mathbf{A}(\alpha\mathbf{u}_1 + \beta\mathbf{u}_2) = \alpha\mathbf{A}\mathbf{u}_1 + \beta\mathbf{A}\mathbf{u}_2 = \alpha\lambda\mathbf{u}_1 + \beta\lambda\mathbf{u}_2 = \lambda(\alpha\mathbf{u}_1 + \beta\mathbf{u}_2) = \lambda\mathbf{u}.$$

Exercise 2 - Variance and Expectation

(a) Given a set of vectors $\{\mathbf{x}_i\}_{i=1}^N$. Show that their empirical mean is equivalent to

$$\hat{\mu} = \arg\min_{\mu} \sum_i \|\mathbf{x}_i - \mu\|^2.$$

(b) There are two equivalent definitons of variance of a random variable. The first one is $\mathbf{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]$ and the second is $\mathbf{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Show that these two definitions actually are equivalent.

Solution

(a) The key idea here is to compute the gradient of the objective function and solve for μ. The gradient is obtained by applying the chain rule resulting in

$$0 = \nabla_\mu \sum_i \|\mathbf{x}_i - \mu\|^2 = -2\sum_i(\mathbf{x}_i - \mu).$$

Now, we solve this for μ to obtain

$$\mu = \frac{1}{N}\sum_i \mathbf{x}_i.$$

(b) Here, we simply need to apply some algebraic manipulations to show that the two definitions are equivalent. We start with the first definition and expand the square:

$$\begin{aligned}
\mathbb{E}[(X - \mathbb{E}[X])^2] &= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}[2X\mathbb{E}[X]] + \mathbb{E}\left[\mathbb{E}[X]^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2
\end{aligned}$$

Exercise 3 - Linear Regression

(a) In the linear regression model with one feature, we have the following model/hypothesis:

$$y = f(x) = wx + b$$

with parameters, w and b, which we wish to find by minimizing the cost:

$$\mathcal{E}(w, b) = \frac{1}{2N} \sum_i ((wx^{(i)} + b) - t^{(i)})^2$$

What are the derivatives $\frac{\partial \mathcal{E}}{\partial w}$ and $\frac{\partial \mathcal{E}}{\partial b}$?

(b) In the linear regression model with many features, we have the following model/hypothesis:

$$y = f(x) = w^\top x + b$$

with parameters, $w = [w_1, w_2, \ldots w_d]^T$ and b, which we wish to find by minimizing the cost:

$$\mathcal{E}(w, b) = \frac{1}{2N} \sum_i ((\mathbf{w}^\top \mathbf{x}^{(i)} + b) - t^{(i)})^2$$

What is the derivative $\frac{\partial \mathcal{E}}{\partial w_j}$ for a weight w_j?

Solution

(a) We obtain the derivative with respect to w directly using the chain rule resulting in

$$\frac{\partial \mathcal{E}}{\partial w} = \frac{1}{N} \sum_i x^{(i)}((wx^{(i)} + b) - t^{(i)})$$

Similarly, the derivative with respect to b is

$$\frac{\partial \mathcal{E}}{\partial b} = \frac{1}{N} \sum_i ((wx^{(i)} + b) - t^{(i)})$$

(b) The derivative with respect to w_j is

$$\frac{\partial \mathcal{E}}{\partial w_j} = \frac{1}{N} \sum_i x_j^{(i)}((\mathbf{w}^\top \mathbf{x}^{(i)} + b) - t^{(i)})$$

Exercise 4 - Gradients and Computation Graphs

(a) Compute the $\frac{\partial \mathcal{L}}{\partial w_j}$ gradient of $\mathcal{L}$ with respect to a w_j in the following computation:

$$\mathcal{L}(y, t) = -t \log(y) - (1 - t) \log(1 - y), \qquad y = \sigma(z), \qquad z = \mathbf{w}^\top \mathbf{x}.$$

(b) Draw the computation graph for the following neural network, showing the relevant scalar quantities. Assume that $\mathbf{y}, \mathbf{h}, \mathbf{x} \in \mathbb{R}^2$

$$\mathcal{L} = \frac{1}{2} \sum_k (y_k - t_k)^2, \qquad y_k = \sum_i w_{ki}^{(2)} h_i + b_k^{(2)}, \qquad h_i = \sigma(z_i), \qquad z_i = \sum_j w_{ij}^{(1)} x_j + b_i^{(1)}.$$

Solution

(a) Applying the chain rule, we have

$$\frac{\partial \mathcal{L}}{\partial w_j} = \frac{\partial \mathcal{L}}{\partial y}\frac{\partial y}{\partial z}\frac{\partial z}{\partial w_j}$$

Looking at each term individually yields

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial y} &= \frac{\partial}{\partial y}[-t\log(y) - (1-t)\log(1-y)] = -\frac{t}{y} + \frac{1-t}{1-y} \\
\frac{\partial y}{\partial z} &= \frac{\partial \sigma(z)}{\partial z} = \sigma(z)(1-\sigma(z)) = y(1-y) \\
\frac{\partial z}{\partial w_j} &= \frac{\partial}{\partial w_j}(w^\top x) = x_j
\end{aligned}$$

Bringing it all together yields:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w_j} &= \left(-\frac{t}{y} + \frac{1-t}{1-y}\right) \cdot y(1-y) \cdot x_j \\
&= (-t + ty + 1 - t - y + ty)x_j \\
&= (y-t)x_j
\end{aligned}$$

(b) The computation graph is given in the figure below.

`nn-compgraph-drawing-eps-converted-to.pdf`

Figure 1: Computation graph for exercise 4 (b)