Pre-Quiz for Machine Learning and Data Science

Yiming Ying, University of Sydney

1. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined, for any $\mathbf{w} = (w_1, w_2)^T \in \mathbb{R}^2$, by

$$f(\mathbf{w}) = \log(1 + e^{-\mathbf{w}^T \mathbf{x}}),$$

where $\mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2$ is a constant column vector.

- Compute the gradient of $f(\mathbf{w})$, i.e., $\nabla f(\mathbf{w}) = (\frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2})^T$.
- Compute the Hessian of f, i.e., $\nabla^2 f(\mathbf{w}) = \begin{bmatrix} \frac{\partial^2 f}{\partial w_1^2} & \frac{\partial^2 f}{\partial w_1 \partial w_2} \\ \frac{\partial^2 f}{\partial w_2 \partial w_1} & \frac{\partial^2 f}{\partial w_2^2} \end{bmatrix}$.
- Show that the Hessian matrix $\nabla^2 f(\mathbf{w})$ is positive semi-definite for any $\mathbf{w}$.

2. Consider the following minimisation problem: $f(\mathbf{w}) = \|\mathbf{b} - A^T\mathbf{w}\|^2 + \|\mathbf{w}\|^2$. Here, $\mathbf{b} \in \mathbb{R}^m$ and matrix $A \in \mathbb{R}^{n\times m}$ (i.e., $n \times m$ matrix) and $\|\cdot\|$ is the standard Euclidean norm, i.e., for any $\mathbf{w} = (w_1, w_2 \ldots, w_n)^T \in \mathbb{R}^n$, $\|\mathbf{w}\| = \sqrt{w_1^2 + w_2^2 + \ldots + w_n^2} = \sqrt{\mathbf{w}^T\mathbf{w}}$. Consider the minimisation problem $\min_{\mathbf{w}\in\mathbb{R}^n} f(\mathbf{w})$.

- Show that $\nabla f(\mathbf{w}) = -2A\mathbf{b} + 2AA^T\mathbf{w} + 2\mathbf{w}$.
- Show that the only minimiser of the above minimisation problem $\min_{\mathbf{w}\in\mathbb{R}^n} f(\mathbf{w})$ is given by $\mathbf{w}^* = (AA^T + \mathbf{I}_n)^{-1}A\mathbf{b}$, where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix.
- Show that $(AA^T + \mathbf{I}_n)^{-1}A = A(A^TA + \mathbf{I}_m)^{-1}$, where $\mathbf{I}_m$ denotes the $m \times m$ identity matrix. Therefore, $\mathbf{w}^* = A(A^TA + \mathbf{I}_m)^{-1}\mathbf{b}$..
- If $n = 10,000$ and $m = 2$, discuss which of the following expressions for $\mathbf{w}^*$ is *computationally easier*:

  $$\mathbf{w}^* = (AA^T + \mathbf{I}_n)^{-1}A\mathbf{b} \text{ or } \mathbf{w}^* = A(A^TA + \mathbf{I}_m)^{-1}\mathbf{b}.$$

  What about the case of $n = 2$ and $m = 10,000$?

3. Suppose that two observations with values $x_1 = 1$ and $x_2 = 2$ are from the random variable which has density $\frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2})$ for some value $\mu \in \mathbb{R}$. What is the maximum likelihood estimate for μ?

4. Consider two parallel hyperplanes in $\mathbb{R}^3$ given by the following two equations

$$ax + by + cz = 1, \qquad ax + by + cz = -1.$$

Show that the distance between these two hyperplanes equals to $\frac{2}{\sqrt{a^2+b^2+c^2}}$.

5. The KL divergence between two distributions with densities p and q is defined to be $KL(q\|p) = \mathbb{E}_q[\log q - \log p]$.

- Show that the KL divergence is non-negative for all distributions q and p.
- Suppose that p and q are two univariate Gaussian distributions with the same deviation σ but with two different means μ_1 and μ_2. Prove that the KL divergence between these two normal distributions equals to $\frac{(\mu_1-\mu_2)^2}{2\sigma^2}$.

Click here for solutions