3M1 Mathematical Methods, 2025

Optimisation

For $x_d \in [0,1]$, one of D independent measurements taking values in the unit interval, the D-dimensional column vector $\mathbf{x} = [x_1, x_2, \cdots, x_d, \cdots, x_D]^\mathsf{T}$ can be formed. The D-dimensional column vector $\mathbf{w} = [w_1, \cdots, w_D]^\mathsf{T} \in \mathbb{R}^D$ represents the D parameters of a single layer neural network that models data $y \in \mathbb{R}$ such that the approximation of the data is denoted as $\hat{y}(\mathbf{x}, \mathbf{w})$ where the measurements are nonlinearly transformed by the function $\phi(\cdot)$,

$$\hat{y}(\mathbf{x}, \mathbf{w}) = \sum_{d=1}^{D} w_d \phi(x_d)$$

The Integrated Squared Error (ISE) is defined as

$$\int_0^1 \int_0^1 \cdots \int_0^1 |\hat{y}(\mathbf{x}, \mathbf{w}) - y|^2 dx_1 dx_2 \cdots dx_D$$

1. By defining the elements of the $D \times D$ matrix $\mathbf{C}$, and $D \times 1$ column vector $\mathbf{b}$, prove that the weights $\mathbf{w}$ that minimise the ISE can be identified by minimising the function of the weights

$$f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{C}\mathbf{w} - \mathbf{w}^\mathsf{T}\mathbf{b}$$

[20%]

The ISE can be written as

$$\begin{aligned} ISE &= \int (y^2 + \hat{y}(\mathbf{x}, \mathbf{w})^2 - 2\hat{y}(\mathbf{x}, \mathbf{w}))d\mathbf{x} \\ &= \int y^2 + \mathbf{w}^\mathsf{T}\mathbf{f}\mathbf{f}^\mathsf{T}\mathbf{w} - 2y\mathbf{w}^\mathsf{T}\mathbf{f} d\mathbf{x} \\ &= y^2 + \mathbf{w}^\mathsf{T} \int \mathbf{f}\mathbf{f}^\mathsf{T} d\mathbf{x} \times \mathbf{w} - 2y\mathbf{w}^\mathsf{T} \int \mathbf{f} d\mathbf{x} \end{aligned}$$

As y not a function of $\mathbf{w}$ then can be dropped and the equivalent function to optimise can be written as

$$\frac{1}{2}\mathbf{w}^\mathsf{T} \int \mathbf{f}\mathbf{f}^\mathsf{T} d\mathbf{x} \times \mathbf{w} - y\mathbf{w}^\mathsf{T} \int \mathbf{f} d\mathbf{x}$$

which both yield same minimum for $\mathbf{w}$. Now $\mathbf{C} = \int \mathbf{f}\mathbf{f}^\mathsf{T} d\mathbf{x}$ with $C_{ij} = \int\int \phi(x_i)\phi(x_j) dx_i dx_j$ and $C_{ii} = \int \phi(x_i)^2 dx_i$ and $b_i = y \int \phi(x_i) dx_i$ 10 marks for the quadratic form, 5 marks for definition of $C_{ij} and C_{ii}$ and 5 marks for b_i.

2. For the case where $D = 2$ and $\phi(\cdot)$ is the identity function such that the neural network approximation is $\hat{y}(\mathbf{x}, \mathbf{w}) = \sum_{d=1}^{2} w_d x_d$, verify that the ISE is globally minimised when the weights take the unique values $w_1 = w_2 = \frac{6y}{7}$. [20%]

$C_{ij} = \int\int x_i x_j dx_i dx_j = \frac{x_i^2}{2}|_0^1 \times \frac{x_j^2}{2}|_0^1 = \frac{1}{4}$ and $C_{ii} = \int x_i^2 dx_i = \frac{x_i^3}{3}|_0^1 = \frac{1}{3}$, and $b_i = y \int x_i dx_i = \frac{y}{2}$ the 2×2 matrix $\mathbf{C} = [1/3\ 1/4; 1/4\ 1/3]$ is the Hessian which is positive and non-zero so is invertible and denotes a global minimum of the ISE and $\mathbf{w}* = \mathbf{C}^{-1}\mathbf{b}$ which yields $w_1 = \frac{6y}{7}$ and $w_2 = \frac{6y}{7}$. 10 marks for defintion of C and b, 10 marks for least squares solution and correct arithmetic plus noting this is unique solution.

3. Consider the addition of a nonlinear term to the function $f(\mathbf{w})$ such that

$$\varphi(\mathbf{w}) = f(\mathbf{w}) + \mathbf{1}^{\mathsf{T}}\mathbf{g}(\mathbf{w})$$

where the D-dimensional vector of ones is denoted as $\mathbf{1} = [1, 1, \cdots, 1]^{\mathsf{T}}$ and $\mathbf{g}(\mathbf{w})$ is an element-wise application of the function $g(\cdot)$ which acts on each component of the vector $\mathbf{w}$ such that $\mathbf{g}(\mathbf{w}) = [g(w_1), g(w_2), \cdots, g(w_D)]^{\mathsf{T}}$. Derive and justify the necessary and sufficient conditions for a point $\mathbf{w}^* \in \mathbb{R}^D$ to be a strong local minimum of the function $\varphi(\mathbf{w})$. [30%]

For a point $\mathbf{w}^*$ to be a local minimum and as there are no constraints so that all directions in $\mathbb{R}^D$ are feasible then the necessary condition is that $\nabla f(\mathbf{w}^*) = 0$ which for the function should be $\mathbf{C}\mathbf{w}^* - \mathbf{b} + \mathbf{g}'(\mathbf{w}^*) = 0$ where $\mathbf{g}'(\mathbf{w}^*)$ is a component wise defined vector with elements $g'(.)$. For an arbitrary vector $\mathbf{d}$ then $\mathbf{d}^{\mathsf{T}}(\mathbf{C} + g''(\mathbf{w}^*))\mathbf{d} > 0$ must be satisfied for $\mathbf{w}^*$ to be a strong local minimum. Where the non-zero elements of the diagonal matrix $g''(\mathbf{w}^*)$ are the second derivatives $g''(w_i^*)$. 15 marks for each neccesary and sufficient.

4. For the specific case where $D = 2$, $\phi(\cdot)$ is the identity function, and the nonlinear term is the quartic polynomial, i.e. $g(w_d) = w_d^4$, assess whether $\varphi(\mathbf{w})$ is convex in $\mathbb{R}^2$ and state the implication on the nature of the point $\mathbf{w}^*$. [30%]

The Hessian is $\mathbf{C} + g''(\mathbf{w}_k)$ which yields a determinant $(C_{11} + g''(w_1))(C_{22} + g''(w_2)) - C_{12}C_{21}$ with $g(w_1)'' = 12w_1^2$ and $g(w_2)'' = 12w_2^2$ that has to be positive or equal to zero for all $\mathbf{w}$ for it to be convex. It is clear that when the quartic function is used for $g(\cdot)$ and the specific values of $\mathbf{C}$ are used then this will be the case, ie. the strict positivity will hold due to the positive value of the squared function for all values of argument. This therefore implies that the minimum is in fact a unique global minimum.

Linear Algebra

1. (a) Consider how 'big' a vector x with $\|x\| = 1$ can become after the multiplication by A.
   - 1-norm: Consider a A applied to a vector x. The result is the sum of the scaled columns of A, with the scale applied to each column being the corresponding entry in x. Subject to $\|x\|_1 = 1$, the maximum norm of the result is equal to the column of A with the maximum 1-norm. When can find equality by selecting an x with 1 in the position that corresponds to the column with the maximum 1-norm. *Therefore the 1-norm of A is equal to the 1-norm of the column of A with the maximum 1-norm.*
   - ∞-norm: In this case $|x_i| \leq 1$, with $|x_i| = 1$ for at least one i (satisfying $\|x\|_\infty = 1$). In this case the resulting vector after applying A is less then or equal the row of A with the greatest 1-norm. We can find equality by setting entries in x to ± 1. *Therefore the ∞-norm of A is equal to the 1-norm of the row of A with the maximum 1-norm.*

   (b) From the definition of the norm,

$$\|Ax\|_\beta \leq \|A\|_{\alpha,\beta}\|x\|_\alpha.$$

For x insert Bx:

$$\|ABx\|_\beta \leq \|A\|_{\gamma,\beta}\|Bx\|_\gamma \leq \|A\|_{\gamma,\beta}\|B\|_{\alpha,\gamma}\|x\|_\alpha.$$

Therefore

$$\|AB\|_{\alpha,\beta} \leq \|A\|_{\gamma,\beta}\|B\|_{\alpha,\gamma}.$$

2. If $Au = \lambda u$, then $BAu = \lambda Bu$. Since $AB = BA$, we have $ABu = \lambda Bu$. This implies that $Bu = 0$ or Bu is an eigenvector of A.

3. (a) If $Pu = \lambda u$, since $P^2 = P$ we have $Pu = P^2u = \lambda Pu = \lambda^2 u$. Therefore $\lambda = 1$ or $\lambda = 0$. Since $P = P^H$, we have $P^2 = PP^H$, and the κ_2 condition number is the ratio of the maximum eigenvalue of PP^H over the smallest. Therefore $\kappa_2(A)$ can be 1 or ∞.

   (b) $(Px)^H(y - Py) = x^H P^H y - x^H P^H Py = x^H P^H y - x^H P^H y = 0$. P generally projects a vector into a subspace. This results says that the 'residual' of the projection, $y - Py$, is orthogonal to all vectors Px, i.e. the sub-space that P projects into.

   (c) $\|Px\|_2^2 = x^H P^H Px = x^H Px \leq \|Px\|_2 \|x\|_2$. Since $\|Px\|_2^2 \leq \|Px\|_2 \|x\|_2$, we have $\|Px\|_2 \leq \|x\|_2$.

4. The $n \times n$ centering matrix has the form:

$$C = I_{n\times n} - \frac{1}{n} 1_{n\times n}.$$

   Computing the product CC,

$$(I_{n\times n} - \frac{1}{n} 1_{n\times n})(I_{n\times n} - \frac{1}{n} 1_{n\times n}) = I_{n\times n} - \frac{2}{n} 1_{n\times n} + \frac{1}{n^2} n_{n\times n} = I_{n\times n} - \frac{1}{n} 1_{n\times n},$$

   proving that it is a projection. It is also orthogonal.

   The only vector in the nullspace is the constant vector, i.e. if $x_i = 1$, $Cx = 0$. Therefore rank(C) is $n - 1$. The centring matrix has one zero eigenvalue and $n - 1$ eigenvalues equal to one.

Stochastic Processes

A simple five-stage manufacturing process can be described by a Markov process. At each time, an item is placed in State 1, the start of the process, just before each transition occurs. Completed items are then kept in State 5, the end of the process. Figure 1 shows the state-diagram for the process.

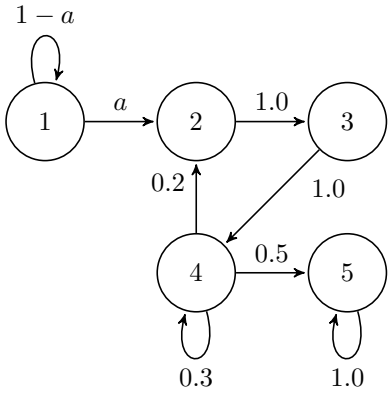

Figure 1: Five-stage manufacturing process described by a Markov process.

1. Write down the transition matrix, $\mathbf{P}$, for this process. What is the stationary distribution for this process? [15%]

The transition matrix is

$$P = \begin{bmatrix} 1-a & a & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0.2 & 0 & 0.3 & 0.5 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

By inspection, the stationary distribution must be

$$\pi = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Unless $a = 0$, in which case $\pi = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \end{bmatrix}$.

2. What is the expected time for an item to first enter State 5? [30%]

Using the notation that q_i is the expected time to first visit State 5 give a start in State i, the equations associated with this are:

$$\begin{aligned} q_1 &= (1-a)(1+q_1) + a(1+q_2) \\ q_2 &= 1 + q_3 \\ q_3 &= 1 + q_4 \\ q_4 &= 0.5 + 0.3(1+q_4) + 0.2(1+q_2) \end{aligned}$$

It is then possible to write:

$$\begin{aligned} q_2 &= 2 + q_4 \\ 0.7q_4 &= 1 + 0.2q_2 \\ \therefore 0.7q_4 &= 1 + 0.2(2+q_4) \implies 0.5q_4 = 1.4 \\ \therefore q_4 &= 2.8 \text{ and } q_2 = 4.8 \end{aligned}$$

Finally, solving for q_1:

$$\begin{aligned} q_1 &= (1-a)(1+q_1) + a(1+q_2) \\ \therefore aq_1 &= 1 + aq_2 \\ \therefore q_1 &= \frac{1}{a} + q_2 = \frac{1}{a} + 4.8 \end{aligned}$$

3. Write an expression in terms of **P** for the probability distribution over the states for a single item n steps after it entered the process. [10%]

The distribution after n steps is

$$\pi^{(n)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \mathbf{P}^n$$

4. The process is run for N steps:

(a) Show that the expected number of items in each state can be expressed as

$$\pi(\mathbf{P}-\mathbf{A})\mathbf{B}^{-1}$$

What are π, $\mathbf{A}$, and $\mathbf{B}$? [30%]

As a new item enters State 1 every step, the distribution is now changed to

$$\pi^{(N)} = \begin{bmatrix}1 & 0 & 0 & 0 & 0\end{bmatrix}\sum_{i=1}^{N}\mathbf{P}^i$$

This is a geometric progression but based on vectors. Using the standard proof for a GP yields:

$$\pi^{(N)} - \pi^{(N)}\mathbf{P} = \begin{bmatrix}1 & 0 & 0 & 0 & 0\end{bmatrix}\left(\sum_{i=1}^{N}\mathbf{P}^i - \sum_{i=1}^{N}\mathbf{P}^{i+1}\right) = \begin{bmatrix}1 & 0 & 0 & 0 & 0\end{bmatrix}\left(\mathbf{P}^1 - \mathbf{P}^{N+1}\right)$$

$$\therefore \pi^{(N)}(\mathbf{I}-\mathbf{P}) = \begin{bmatrix}1 & 0 & 0 & 0 & 0\end{bmatrix}(\mathbf{P}-\mathbf{P}^{N+1})$$

$$\therefore \pi^{(N)} = \begin{bmatrix}1 & 0 & 0 & 0 & 0\end{bmatrix}(\mathbf{P}-\mathbf{P}^{N+1})(\mathbf{I}-\mathbf{P})^{-1}$$

Hence

$$\pi = \begin{bmatrix}1 & 0 & 0 & 0 & 0\end{bmatrix}$$
$$\mathbf{A} = \mathbf{P}^{N+1}$$
$$\mathbf{B} = \mathbf{I}-\mathbf{P}$$

(b) How does the value of "a" influence the expected number of items in eawch of the states? [15%]

"a" influences the probability of a single item entering the process. Thus the number of items stored in State 1 is determined by this. Again, this can be written as a geometric progression. As N becomes large the system will converge on a steady state, as once an item enters the process the process it is not influenced by "a".