Problem 1: Derivation of the Normal Equations

We consider the polynomial regression model defined by:

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j (x_n)^j$$

The Sum-of-Squares Error function is:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \Big[y(x_n, \mathbf{w}) - t_n \Big]^2$$
$$\Rightarrow E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big]^2$$

To minimize the error, we take the partial derivative with respect to a specific weight w_i:

$$\frac{\partial E}{\partial w_i} = \frac{1}{2} \sum_{n=1}^{N} 2 \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big] \frac{\partial}{\partial w_i} \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big]$$
$$= \sum_{n=1}^{N} \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big] \sum_{j=0}^{M} (x_n)^j \frac{\partial w_j}{\partial w_i}$$

Using the property that weights are independent, $\frac{\partial w_j}{\partial w_i} = \delta_{ij}$ (the Kronecker delta), which is 1 if $j = i$ and 0 otherwise:

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N} \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big] (x_n)^i$$

Setting the gradient to zero to find the minimum:

$$\frac{\partial E}{\partial w_i} = 0$$
$$\Rightarrow \sum_{n=1}^{N} \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big] (x_n)^i = 0$$
$$\Rightarrow \sum_{n=1}^{N} \sum_{j=0}^{M} w_j (x_n)^j (x_n)^i - \sum_{n=1}^{N} t_n (x_n)^i = 0$$
$$\Rightarrow \sum_{n=1}^{N} \sum_{j=0}^{M} w_j (x_n)^{j+i} = \sum_{n=1}^{N} t_n (x_n)^i$$
$$\Rightarrow \sum_{j=0}^{M} w_j \left(\sum_{n=1}^{N} (x_n)^{j+i} \right) = \sum_{n=1}^{N} t_n (x_n)^i$$

We define the matrix elements A_{ij} and vector elements T_i:

$$\sum_{j=0}^{M} A_{ij} w_j = T_i \quad \text{where} \quad A_{ij} = \sum_{n=1}^{N} (x_n)^{j+i}, \quad T_i = \sum_{n=1}^{N} t_n (x_n)^i$$

In matrix notation, this linear system is:

$$\begin{aligned} \mathbf{A}\mathbf{w} &= \mathbf{T} \\ \Rightarrow \mathbf{w} &= \mathbf{A}^{-1}\mathbf{T} \end{aligned}$$

Problem 2: Regularized Least Squares

We introduce an L_2 regularization term (weight decay) to the error function:

$$\begin{aligned} \tilde{E}(\mathbf{w}) &= \frac{1}{2} \sum_{n=1}^{N} \Big[y(x_n, \mathbf{w}) - t_n \Big]^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\ \Rightarrow \tilde{E}(\mathbf{w}) &= \frac{1}{2} \sum_{n=1}^{N} \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big]^2 + \frac{\lambda}{2} \sum_{j=0}^{M} w_j^2 \end{aligned}$$

Differentiating with respect to w_i:

$$\begin{aligned} \frac{\partial \tilde{E}}{\partial w_i} &= \frac{\partial}{\partial w_i} \left(\frac{1}{2} \sum_{n=1}^{N} \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big]^2 \right) + \frac{\partial}{\partial w_i} \left(\frac{\lambda}{2} \sum_{j=0}^{M} w_j^2 \right) \\ &= \sum_{n=1}^{N} \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big] (x_n)^i + \frac{\lambda}{2} (2 w_i) \\ &= \sum_{n=1}^{N} \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big] (x_n)^i + \lambda w_i \end{aligned}$$

Setting the gradient to zero:

$$\begin{aligned} &\frac{\partial \tilde{E}}{\partial w_i} = 0 \\ &\Rightarrow \sum_{n=1}^{N} \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big] (x_n)^i + \lambda w_i = 0 \\ &\Rightarrow \sum_{n=1}^{N} \sum_{j=0}^{M} w_j (x_n)^{j+i} - \sum_{n=1}^{N} t_n (x_n)^i + \lambda w_i = 0 \\ &\Rightarrow \sum_{j=0}^{M} w_j \left(\sum_{n=1}^{N} (x_n)^{j+i} \right) + \lambda w_i = \sum_{n=1}^{N} t_n (x_n)^i \end{aligned}$$

Using the same definitions for A_{ij} and T_i as in Problem 1:

$$\sum_{j=0}^{M} A_{ij} w_j + \lambda w_i = T_i$$

In matrix notation, λw_i corresponds to adding λ to the diagonal elements of $\mathbf{A}$:

$$\begin{aligned}
&\mathbf{A}\mathbf{w} + \lambda \mathbf{I}\mathbf{w} = \mathbf{T} \\
&\Rightarrow (\mathbf{A} + \lambda \mathbf{I})\mathbf{w} = \mathbf{T} \\
&\Rightarrow \mathbf{w} = (\mathbf{A} + \lambda \mathbf{I})^{-1}\mathbf{T}
\end{aligned}$$

Problem 4: Calculus of Variations for Optimal Prediction

We seek to minimize the expected loss functional $E[L]$ with respect to the function $\mathbf{y}(\mathbf{x})$:

$$\begin{aligned}
E[L] &= \iint \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t})\, d\mathbf{x}\, d\mathbf{t} \\
\text{Expanding the norm:} \quad \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 &= \sum_{k=1}^{M} (\mathbf{y}_k(\mathbf{x}) - \mathbf{t}_k)^2 \\
\Rightarrow E[L] &= \iint \sum_{k=1}^{M} (\mathbf{y}_k(\mathbf{x}) - \mathbf{t}_k)^2 p(\mathbf{x}, \mathbf{t})\, d\mathbf{x}\, d\mathbf{t}
\end{aligned}$$

We take the functional derivative $\delta E[L]/\delta \mathbf{y}_j(\mathbf{x})$. The derivative enters the integral:

$$\frac{\delta E[L]}{\delta \mathbf{y}_j(\mathbf{x})} = \iint \frac{\partial}{\partial \mathbf{y}_j(\mathbf{x})} \left[\sum_{k=1}^{M} (\mathbf{y}_k(\mathbf{x}) - \mathbf{t}_k)^2 \right] p(\mathbf{x}, \mathbf{t})\, d\mathbf{x}\, d\mathbf{t}$$

The derivative of the squared term is:

$$\frac{\partial}{\partial \mathbf{y}_j(\mathbf{x})} (\mathbf{y}_j(\mathbf{x}) - \mathbf{t}_j)^2 = 2(\mathbf{y}_j(\mathbf{x}) - \mathbf{t}_j)$$

Thus:

$$\frac{\delta E[L]}{\delta \mathbf{y}_j(\mathbf{x})} = \iint 2(\mathbf{y}_j(\mathbf{x}) - \mathbf{t}_j) p(\mathbf{x}, \mathbf{t})\, d\mathbf{x}\, d\mathbf{t}$$

Because we vary $\mathbf{y}(\mathbf{x})$ at a specific point $\mathbf{x}$, we can remove the integral over $d\mathbf{x}$ (effectively utilizing the Dirac delta property of functional derivatives):

$$\frac{\delta E[L]}{\delta \mathbf{y}_j(\mathbf{x})} = 2 \int (\mathbf{y}_j(\mathbf{x}) - \mathbf{t}_j) p(\mathbf{x}, \mathbf{t})\, d\mathbf{t}$$

Setting the derivative to zero for optimality:

$$\begin{aligned}
\int (\mathbf{y}_j(\mathbf{x}) - \mathbf{t}_j) p(\mathbf{x}, \mathbf{t})\, d\mathbf{t} &= 0 \\
\mathbf{y}_j(\mathbf{x}) \int p(\mathbf{x}, \mathbf{t})\, d\mathbf{t} - \int \mathbf{t}_j p(\mathbf{x}, \mathbf{t})\, d\mathbf{t} &= 0
\end{aligned}$$

Recognizing that $\int p(\mathbf{x}, \mathbf{t})\, d\mathbf{t} = p(\mathbf{x})$:

$$\mathbf{y}_j(\mathbf{x})p(\mathbf{x}) = \int \mathbf{t}_j p(\mathbf{x}, \mathbf{t})\, d\mathbf{t}$$

Solving for $\mathbf{y}_j(\mathbf{x})$ and using $p(\mathbf{x}, \mathbf{t}) = p(\mathbf{t}|\mathbf{x})p(\mathbf{x})$:

$$\begin{aligned}
\mathbf{y}_j(\mathbf{x}) &= \frac{\int \mathbf{t}_j p(\mathbf{x}, \mathbf{t})\, d\mathbf{t}}{p(\mathbf{x})} \\
&= \int \mathbf{t}_j \frac{p(\mathbf{x}, \mathbf{t})}{p(\mathbf{x})}\, d\mathbf{t} \\
&= \int \mathbf{t}_j p(\mathbf{t} \mid \mathbf{x})\, d\mathbf{t} \\
&= \mathbb{E}[\mathbf{t}_j \mid \mathbf{x}]
\end{aligned}$$

Thus, the optimal prediction is the conditional expectation of the target.

Problem 5: The Binomial Distribution

5.1 Proof of Pascal's Identity

We wish to prove:

$$\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}$$

Expanding using factorials:

$$\binom{N}{m} + \binom{N}{m-1} = \frac{N!}{m!(N-m)!} + \frac{N!}{(m-1)!(N-m+1)!}$$

Find a common denominator. Multiply the first term by $(N - m + 1)$ and the second by m:

$$\begin{aligned}
&= \frac{N!(N-m+1)}{m!(N-m)!(N-m+1)} + \frac{N!(m)}{m(m-1)!(N-m+1)!} \\
&= \frac{N!(N-m+1)}{m!(N-m+1)!} + \frac{N!(m)}{m!(N-m+1)!} \\
&= \frac{N!\big[(N-m+1)+m\big]}{m!(N-m+1)!} \\
&= \frac{N!(N+1)}{m!(N+1-m)!} \\
&= \frac{(N+1)!}{m!((N+1)-m)!} \\
&= \binom{N+1}{m}
\end{aligned}$$

5.2 Proof of Binomial Theorem by Induction

We prove $(1+x)^N = \sum_{m=0}^{N} \binom{N}{m} x^m$ for integer $N \geq 0$.

Base Case ($N = 0$):

$$\text{LHS: } (1+x)^0 = 1 \qquad \text{RHS: } \sum_{m=0}^{0} \binom{0}{m} x^m = \binom{0}{0} x^0 = 1$$

The base case holds.

Inductive Step: Assume the hypothesis holds for N. Consider $N+1$:

$$\begin{aligned}
(1+x)^{N+1} &= (1+x)(1+x)^N \\
&= (1+x)\sum_{m=0}^{N} \binom{N}{m} x^m \quad \text{(by hypothesis)} \\
&= \sum_{m=0}^{N} \binom{N}{m} x^m + \sum_{m=0}^{N} \binom{N}{m} x^{m+1}
\end{aligned}$$

We shift the index of the second summation. Let $k = m+1$. When $m = 0, k = 1$. When $m = N, k = N+1$.

$$\sum_{m=0}^{N} \binom{N}{m} x^{m+1} = \sum_{k=1}^{N+1} \binom{N}{k-1} x^k$$

Renaming k back to m for consistency:

$$\begin{aligned}
(1+x)^{N+1} &= \binom{N}{0} x^0 + \sum_{m=1}^{N} \binom{N}{m} x^m + \sum_{m=1}^{N} \binom{N}{m-1} x^m + \binom{N}{N} x^{N+1} \\
&= \binom{N}{0} x^0 + \sum_{m=1}^{N} \left[\binom{N}{m} + \binom{N}{m-1} \right] x^m + \binom{N}{N} x^{N+1}
\end{aligned}$$

Using Pascal's Identity, and noting $\binom{N}{0} = 1 = \binom{N+1}{0}$ and $\binom{N}{N} = 1 = \binom{N+1}{N+1}$:

$$\begin{aligned}
(1+x)^{N+1} &= \binom{N+1}{0} x^0 + \sum_{m=1}^{N} \binom{N+1}{m} x^m + \binom{N+1}{N+1} x^{N+1} \\
&= \sum_{m=0}^{N+1} \binom{N+1}{m} x^m
\end{aligned}$$

This completes the proof by induction.

5.3 Normalization of the Binomial Distribution

The Binomial distribution is given by:

$$\mathrm{Bin}(m \mid N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

We verify it sums to 1:

$$\sum_{m=0}^{N} \mathrm{Bin}(m \mid N, \mu) = \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

Using the Binomial Theorem $(a+b)^N = \sum_{m=0}^{N} \binom{N}{m} a^m b^{N-m}$ with $a = \mu$ and $b = 1-\mu$:

$$\begin{aligned}\sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} &= (\mu + (1-\mu))^N \\ &= (1)^N \\ &= 1\end{aligned}$$