Bayesian Linear Regression — Practice Questions

CM52054: Foundational Machine Learning
Practice set with fully worked answers

Conceptual Questions

1) **Probabilistic model for linear regression.**
What is the probabilistic model assumed for linear regression?

Answer: For each data point (x_i, y_i), the model is

$$y_i = w^\top x_i + \varepsilon_i,$$

where the noise term ε_i is i.i.d. Gaussian:

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Equivalently,

$$p(y_i \mid x_i, w) \sim \mathcal{N}(w^\top x_i, \sigma^2).$$

2) **Meaning of i.i.d. noise.**
What does "i.i.d. noise" mean in this context?

Answer: "i.i.d." stands for independent and identically distributed. In this context it means:

- *Independent*: the noise terms $\varepsilon_1, \ldots, \varepsilon_N$ are statistically independent across data points.
- *Identically distributed*: each noise term follows the same distribution $\mathcal{N}(0, \sigma^2)$.

So each observation is corrupted by an independent draw from the same Gaussian noise distribution.

3) **Likelihood for the whole dataset.**
Write down the likelihood $p(y \mid X, w)$ for the whole dataset under the Gaussian noise model.

Answer: Let

- $X \in \mathbb{R}^{N \times M}$ be the data matrix whose i-th row is $x_i^\top$,
- $y \in \mathbb{R}^N$ be the vector of labels.

Then

$$p(y \mid X, w) = \prod_{i=1}^{N} p(y_i \mid x_i, w) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - w^\top x_i)^2}{2\sigma^2}\right).$$

In compact form:

$$p(y \mid X, w) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{\|Xw - y\|^2}{2\sigma^2}\right).$$

4) **Optimisation problem from ML estimation.**
What optimisation problem does Maximum Likelihood (ML) estimation lead to in this setting?

Answer: ML chooses

$$w^* = \arg\max_w p(y \mid X, w).$$

Maximising the Gaussian likelihood is equivalent to minimising the squared error:

$$w^* = \arg\min_{w \in \mathbb{R}^M} \|Xw - y\|^2.$$

So ML estimation under i.i.d. Gaussian noise is equivalent to standard least squares.

5) **Difference between ML and MAP.**
What is the difference between ML and MAP estimation?

Answer:

- **ML (Maximum Likelihood):**

$$w^*_{\text{ML}} = \arg\max_w p(y \mid X, w).$$

  It only uses the likelihood (data fit), and ignores any prior information on w.
- **MAP (Maximum A Posteriori):**

$$w^*_{\text{MAP}} = \arg\max_w p(w \mid X, y) = \arg\max_w p(y \mid X, w)\, p(w).$$

  It maximises the posterior by combining the likelihood with a prior $p(w)$. This introduces regularisation / a bias toward parameters that are more probable a priori.

6) **Effect of a Gaussian prior.**
If we set a Gaussian prior $p(w) = \mathcal{N}(0, I)$, how does this affect the solution compared to ML?

Answer: With Gaussian prior $p(w) \propto \exp(-\|w\|^2/2)$, the MAP objective becomes

$$w^*_{\text{MAP}} = \arg\max_w p(y \mid X, w)\, p(w) = \arg\min_w (\|Xw - y\|^2 + \sigma^2 \|w\|^2).$$

So compared to ML (which minimises $\|Xw - y\|^2$ only), the MAP solution adds an additional regularised term $\sigma^2\|w\|^2$. This biases the weights toward zero and can reduce overfitting.

Short Derivations and Calculations

7) **Normal equations for the ML estimator.**
Starting from

$$w^* = \arg\min_w \|Xw - y\|^2,$$

derive the condition satisfied by w^* (the "normal equations").

Answer: Define the objective:

$$J(w) = \|Xw - y\|^2 = (Xw - y)^\top (Xw - y).$$

Take the gradient with respect to w:

$$\nabla_w J(w) = 2X^\top (Xw - y).$$

Set to zero at optimum:

$$2X^\top(Xw^* - y) = 0 \quad \Rightarrow \quad X^\top X w^* = X^\top y.$$

These are the normal equations. If $X^\top X$ is invertible, then

$$w^* = (X^\top X)^{-1} X^\top y.$$

8) **MAP with Gaussian prior**
Assume

$$\text{Likelihood:} \qquad p(y \mid X, w) \propto \exp\Big(-\frac{1}{2\sigma^2}\|Xw - y\|^2\Big),$$

$$\text{Prior:} \qquad p(w) \propto \exp\Big(-\frac{1}{2}\|w\|^2\Big).$$

Show that maximising the posterior is equivalent to minimising $\|Xw - y\|^2 + \sigma^2\|w\|^2$.

Answer: Posterior up to proportionality:

$$p(w \mid X, y) \propto p(y \mid X, w)p(w) \propto \exp\left(-\frac{1}{2\sigma^2}\|Xw - y\|^2\right) \exp\left(-\frac{1}{2}\|w\|^2\right).$$

Combine exponents:

$$\log p(w \mid X, y) = \text{const} - \frac{1}{2\sigma^2}\|Xw - y\|^2 - \frac{1}{2}\|w\|^2.$$

Maximising the log posterior is equivalent to minimising

$$\frac{1}{\sigma^2}\|Xw - y\|^2 + \|w\|^2.$$

Multiplying by σ^2 (a positive constant that does not change the minimiser):

$$w^*_{\text{MAP}} = \arg\min_w \left[\|Xw - y\|^2 + \sigma^2\|w\|^2\right].$$

Thus, MAP with Gaussian prior is equivalent to an ℓ_2-regularised least-squares problem.

9) **Closed-form solution of the MAP estimator.**
Starting from the MAP objective

$$J_{\text{MAP}}(w) = \|Xw - y\|^2 + \sigma^2\|w\|^2,$$

derive w^*_{MAP}.

Answer: Let

$$J_{\text{MAP}}(w) = (Xw - y)^\top(Xw - y) + \sigma^2 w^\top w.$$

Differentiate w.r.t. w:

$$\nabla_w J_{\text{MAP}}(w) = 2X^\top(Xw - y) + 2\sigma^2 w.$$

Set gradient to zero:

$$2X^\top(Xw^* - y) + 2\sigma^2 w^* = 0.$$

Divide by 2:

$$X^\top X w^* - X^\top y + \sigma^2 w^* = 0.$$

Rearrange:

$$(X^\top X + \sigma^2 I)w^* = X^\top y.$$

Assuming $X^\top X + \sigma^2 I$ is invertible:

$$w^*_{\text{MAP}} = (X^\top X + \sigma^2 I)^{-1} X^\top y.$$

10) **Comparing ML and MAP solutions.**
Write down the ML and MAP solutions side by side. How are they related?

Answer:

- ML solution:
$$w^*_{\text{ML}} = (X^\top X)^{-1} X^\top y,$$
assuming $X^\top X$ is invertible.
- MAP solution (with Gaussian prior $\mathcal{N}(0, I)$):
$$w^*_{\text{MAP}} = (X^\top X + \sigma^2 I)^{-1} X^\top y.$$

Relation: MAP adds a positive ridge term $\sigma^2 I$ to $X^\top X$, which

- makes the matrix more stable/invertible (even if $X^\top X$ is nearly singular),
- shrinks parameters toward zero, reducing variance and potential overfitting.