Problem Set 3 - Machine Learning Theories

May 21, 2025

Problem 1 (Bayesian Model Selection)
Consider the problem of selecting between two different regression models, M_0 and M_1, to explain a dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$. We are using a Bayesian approach.

- Model M_0 is a simple model with parameters $\mathbf{w}_0$.
- Model M_1 is a more complex model with parameters $\mathbf{w}_1$.

After performing Bayesian inference, the log marginal likelihood (also known as log evidence) for each model is calculated as:

$$\ln p(D|M_0) = -25.6$$

$$\ln p(D|M_1) = -23.1$$

Assume we have no prior preference for either model, i.e., $p(M_0) = p(M_1) = 0.5$.

(1) Explain how the marginal likelihood $p(D|M_k)$ is used in Bayesian model selection. Why is it a principled way to compare models?

(2) The Bayes factor B_{10} is defined as the ratio of marginal likelihoods: $B_{10} = \frac{p(D|M_1)}{p(D|M_0)}$. Calculate B_{10} using the provided log marginal likelihoods. Based on this Bayes factor, which model is preferred by the data? Explain how the magnitude of the Bayes factor can be interpreted to assess the strength of evidence for the preferred model.

(3) If the prior probabilities were $p(M_0) = 0.8$ and $p(M_1) = 0.2$, would this change which model has a higher posterior probability $p(M_k|D)$? Briefly explain your reasoning without detailed recalculation of posterior probabilities, but by considering the components of Bayes' theorem for model posteriors.

Problem 2 (Bayesian Information Criterion and Model Selection)
In this problem, you will conduct a numerical experiment in model selection using the Bayesian Information Criterion (BIC). The goal is to identify which of several candidate covariance structures best explains synthetic data generated from a multivariate Gaussian distribution.

Use the provided notebook, `Homework3-2.ipynb`, and follow the steps below:

Data generation
Using the notebook, generate datasets $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of size $N = 150$, where each observation $\mathbf{x}_i \in \mathbb{R}^2$ come from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma})$. The data can be simulated under one of three different settings for the covariance matrix $\mathbf{\Sigma}$.

Model candidates
You will fit three different models to the data and compare them using BIC:

- Model M_1 (Spherical Covariance): Assumes $\mathbf{\Sigma} = \sigma^2\mathbf{I}$.
- Model M_2 (Diagonal Covariance): Assumes $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$.
- Model M_3 (Full Covariance): Assumes a general symmetric positive definite $\mathbf{\Sigma}$.

(1) State the formula for the Bayesian Information Criterion (BIC) and explain the role of each term in the context of model selection.

(2) Perform three separate experiments by changing the value of `true_model_type` variable for data generation in `Homework3-2.ipynb`:

- 'spherical' (Experiment A)
- 'diagonal' (Experiment B)
- 'full' (Experiment C)

For each experiment, fit all three models M_1, M_2, M_3 to the generated data. For each of the three *fitted* models (M_1, M_2, M_3), record the number of parameters (k), the maximized log-likelihood $(\ln p(D|\hat{\theta}, M))$, and the BIC score (using $N = 150$ and $\ln(150) \approx 5.01$, or verify the $\ln(N)$ value used in the notebook). Then identify which model is selected by BIC in each of the three experiments. Present these collected results clearly, for example, by using a table.

(3) Based on your experimental results and the BIC-selected models identified in question (2), for each of your three experiments, explain why BIC preferred the specific model it selected in that particular experiment. Your explanation should focus on how the interplay of that model's maximized log-likelihood, its number of parameters (k), and the sample size $(N = 150)$ contributed to its BIC score and subsequent selection over the other candidate models.

Problem 3 (Bayesian Logistic Regression)
Consider a binary classification problem with input features $\mathbf{x} \in \mathbb{R}^d$ and binary labels $y \in \{0, 1\}$. In Bayesian logistic regression, the probability of $y = 1$ given $\mathbf{x}$ and model parameters (weights) $\mathbf{w} \in \mathbb{R}^d$ is modeled using the sigmoid function $\sigma(z) = (1 + e^{-z})^{-1}$:

$$p(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x})$$

We place a Gaussian prior on the weights $\mathbf{w}$:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

where $\mathbf{m}_0$ is the prior mean vector and $\mathbf{S}_0$ is the prior covariance matrix. Suppose we have a dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ of N independent and identically distributed observations.

(1) For a single data point $(\mathbf{x}_i, y_i)$, write down the likelihood $p(y_i|\mathbf{x}_i, \mathbf{w})$. Then, write down the log-likelihood for the entire dataset D, denoted as $\ln p(D|\mathbf{w})$.

(2) Using Bayes' theorem, write down an expression for the log-posterior distribution $\ln p(\mathbf{w}|D)$ up to an additive constant that does not depend on $\mathbf{w}$.

(3) Explain why the posterior distribution $p(\mathbf{w}|D)$ for Bayesian logistic regression is not analytically tractable.

(4) Derive an expression for the gradient of the log-posterior distribution with respect to the weights $\mathbf{w}$, i.e., $\nabla_{\mathbf{w}} \ln p(\mathbf{w}|D)$. Then, write down the update rule for the gradient ascent algorithm that uses this gradient to find the Maximum A Posteriori (MAP) estimate of $\mathbf{w}$.

Problem 4 (Marginal Likelihood Calculation and Properties)
Consider a scenario where we observe N data points $D = \{x_1, x_2, \ldots, x_N\}$ drawn independently from a Gaussian distribution with an unknown mean μ and a known variance σ^2. The likelihood for a single data point x_i is $p(x_i|\mu, \sigma^2) = \mathcal{N}(x_i|\mu, \sigma^2)$. We place a Gaussian prior distribution on the unknown mean μ:

$$p(\mu|\mu_0, \sigma_0^2) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$$

where μ_0 and σ_0^2 are the known mean and variance of the prior distribution, respectively.

(1) Define the marginal likelihood $p(D|\sigma^2, \mu_0, \sigma_0^2)$ in the context of Bayesian inference. What does this quantity represent and why is it important?

(2) Derive an analytical expression for the log marginal likelihood, $\ln p(D|\sigma^2, \mu_0, \sigma_0^2)$, for the given dataset D, likelihood, and prior. This involves integrating out the unknown mean μ. (Hint: The product of two Gaussian functions is proportional to another Gaussian function. You may need to complete the square in the exponent.)

(3) Consider the case where the prior variance σ_0^2 becomes very large ($\sigma_0^2 \to \infty$), representing a very uninformative prior.

   (a) Analyze the behavior of the log marginal likelihood, $\ln p(D|\sigma^2, \mu_0, \sigma_0^2)$, as $\sigma_0^2 \to \infty$. Does it approach a finite value, $+\infty$, or $-\infty$? Justify your answer based on your derived expression.

   (b) Briefly discuss the implications of this result for model comparison if this model (with a very large σ_0^2) were being compared to another model with a more informative prior or a different structure. How does a very uninformative prior affect the model's evidence?

Problem 5 (Laplace Approximation)

Laplace approximation is a technique used to approximate a probability distribution, often when the distribution is intractable to normalize or compute moments from directly. It approximates a unimodal distribution with a Gaussian distribution centered at its mode.

Consider a posterior distribution for a single parameter θ which is proportional to $p(\theta|D) \propto \exp(L(\theta))$, where $L(\theta) = \ln p(D|\theta) + \ln p(\theta)$ is the unnormalized log-posterior.

(1) Briefly explain the main idea behind Laplace's approximation. What are the key steps involved in approximating $p(\theta|D)$ with a Gaussian distribution $q(\theta) = \mathcal{N}(\theta|\theta_{\text{MAP}}, \sigma^2_{\text{Lap}})$?

(2) Suppose the unnormalized log-posterior for a parameter θ is given by:

$$L(\theta) = -\frac{1}{20}(\theta - 2)^4 - (\theta - 3)^2 + \text{const}$$

Let θ_{MAP} be the mode of this distribution (you do not need to find it explicitly for this part). Let $A = -\frac{d^2L(\theta)}{d\theta^2}\Big|_{\theta=\theta_{\text{MAP}}}$. Assuming $A > 0$, how is A related to the variance σ^2_{Lap} of the approximating Gaussian distribution in Laplace's method?

(3) Consider a specific unnormalized log-posterior:

$$L(\theta) = \ln(\cos(\theta)) - \frac{1}{2}\theta^2 \quad \text{for } \theta \in (-\pi/2, \pi/2)$$

Find the mode θ_{MAP} of this distribution (i.e., the value of θ that maximizes $L(\theta)$). Then, calculate the variance σ^2_{Lap} of the Gaussian distribution that approximates $p(\theta|D) \propto e^{L(\theta)}$ around this mode using Laplace approximation.

Problem 6 (Bayesian Logistic Regression)
Consider approximating the variance of the predictive probability $\hat{y} = P(y = 1|\mathbf{x}, \mathbf{w})$ in Bayesian logistic regression.

(1) Show that the following identity holds:

$$\int_{-\infty}^{\infty} \Phi(\lambda a)^2 q_{\mathbf{x}}(a) da = \Phi_2 \left(\frac{\lambda \mu(\mathbf{x})}{\sqrt{1 + \lambda^2 \sigma^2(\mathbf{x})}}, \frac{\lambda \mu(\mathbf{x})}{\sqrt{1 + \lambda^2 \sigma^2(\mathbf{x})}}, \frac{\lambda^2 \sigma^2(\mathbf{x})}{1 + \lambda^2 \sigma^2(\mathbf{x})} \right), \tag{1}$$

where $q_{\mathbf{x}}(a) = N(a; \mu(\mathbf{x}), \sigma^2(\mathbf{x}))$ is a Gaussian density, $\Phi(\cdot)$ denotes the probit function (or the standard normal cumulative distribution function (CDF)), and $\Phi_2(t_1; t_2; \rho)$ denotes the CDF of bivariate standard normal with correlation ρ, i.e., $\Phi_2(t_1; t_2; \rho) = P(x_1 \leq t_1, x_2 \leq t_2)$, where $(x_1, x_2) \sim N(0, \Sigma(\rho))$, $\Sigma(\rho) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

(2) Using the result in (1), derive an approximation for the variance of the predictive probability $\hat{y}$ in Bayesian logistic regression, which is defined as follows:

$$\text{Var}(\hat{y}|\mathbf{x}, \mathbf{y}, X) = \mathbb{E}_{\mathbf{w}}[P(y = 1|\mathbf{x}, \mathbf{w})^2] - \mathbb{E}_{\mathbf{w}}[P(y = 1|\mathbf{x}, \mathbf{w})]^2,$$

where the expectation is taken with respect to the posterior distribution $p(\mathbf{w}|\mathbf{y}, X)$.
Approximate the logistic sigmoid function using the probit function:

$$P(y = 1|\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-a)} \approx \Phi(\lambda a) = \int_{-\infty}^{\lambda a} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx,$$

where $a = \mathbf{w}^\top \mathbf{x}$ and $\lambda = \sqrt{\frac{\pi}{8}}$.
Approximate the posterior distribution of $\mathbf{w}$ using a Gaussian:

$$p(\mathbf{w}|\mathbf{y}, X) \approx N(\mathbf{w}_{MAP}, S_N).$$

Express your answer using the CDF of bivariate normal in (1) and the probit function $\Phi(\cdot)$.