CS329 Machine Learning

Midterm Exam - 2023 Fall

Exam Setter: Prof. Qi Hao
haoq@sustech.edu.cn

Layout: Site Fan
fanst2021@mail.sustech.edu.cn

Question 1 Least Square

a) Consider $Y = AX + V$ and $V \sim \mathcal{N}(\mathbf{v}|\mathbf{0}, Q)$, what is the least square solution of X?

b) If there is a constraint $b^{\mathrm{T}} X = c$, what is the optimal solution of X?

c) If there is an *additional* constraint of $X^{\mathrm{T}} X = d$, in addition to the constraint in b), what is the optimal solution of X?

d) If both A and X are unknown, how to solve A and X alternatively by using two constraints of $X^{\mathrm{T}} X = d$ and $\mathrm{Trace}(A^{\mathrm{T}} A) = e$?

Question 2 Linear Gaussian System

Consider $Y = AX + V$, where X and V are Gaussian, $X \sim \mathcal{N}(\mathbf{x}|\mathbf{m}_0, \mathbf{\Sigma}_0)$, $V \sim \mathcal{N}(\mathbf{v}|\mathbf{0}, \beta^{-1}\mathbf{I})$.

Calculate the followings:

- conditional distribution $p(Y|X)$
- joint distribution $p(Y, X)$
- marginal distribution $p(Y)$
- posterior distribution $p(X|Y = \mathbf{y}, \beta, \mathbf{m}_0, \mathbf{\Sigma}_0)$
- posterior predictive distribution $p\left(\hat{Y}|Y = \mathbf{y}, \beta, \mathbf{m}_0, \mathbf{\Sigma}_0\right)$
- prior predictive distribution $p(Y|\beta, \mathbf{m}_0, \mathbf{\Sigma}_0)$

Question 3 Linear Regression

Consider $y = \mathbf{w}\phi(x) + v$, where v is Gaussian, *i.e.*, $v \sim \mathcal{N}(v|0, \beta^{-1})$, and $\mathbf{w}$ has a Gaussian *priori*, *i.e.*, $\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \alpha^{-1}\mathbf{I})$.

Assume that $\phi(x)$ is known, please derive

- posterior distribution $p(\mathbf{w}|D, \beta, \mathbf{m}_0, \alpha)$
- posterior predictive distribution $p(\hat{y}|\hat{x}, D, \beta, \mathbf{m}_0, \alpha)$
- prior predictive distribution $p(D|\beta, \mathbf{m}_0, \alpha)$

where $D = \{\phi_n, y_n\}, n = 1, ..., N$ is the training dataset and $\phi_n = \phi(x_n)$

Question 4 Logistic Regression

Consider a two-class classification problem with the logistic sigmoid function, $y = \sigma(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}))$, for a given dataset $D = \{\phi_n, t_n\}$, where $t_n \in \{0, 1\}$, $\phi_n = \phi(\mathbf{x}_n), n = 1, ..., N$.

The likelihood function is given by

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{1-t_n}$$

where $\mathbf{w}$ has a Gaussian *priori*, *i.e.*, $\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \alpha^{-1}\mathbf{I})$.

Please derive the followings:

- posterior distribution $p(\mathbf{w}|D, \mathbf{m}_0, \alpha)$
- posterior predictive distribution $p(t|x, D, \mathbf{m}_0, \alpha)$

- prior predictive distribution $p(D|\mathbf{m}_0, \alpha)$

Hint: use Delta approximation and Laplace approximation properly.

Question 5 Neural Network

Consider a two-layer neural network described by the following equations:

$$\begin{aligned} a_1 &= w^{(1)}x, \quad a_1 &= w^{(2)}z \\ z &= h(a_1), \quad y &= \sigma(a_2) \end{aligned}$$

where $\mathbf{x}$ and y are the input and output of the neural network, $h(\cdot)$ is a nonlinear function, and $\sigma(\cdot)$ is the sigmoid function.

1. Please derive the following gradients: $\frac{\partial y}{\partial w^{(1)}}$, $\frac{\partial y}{\partial w^{(2)}}$, $\frac{\partial y}{\partial a_1}$, $\frac{\partial y}{\partial a_2}$ and $\frac{\partial y}{\partial x}$.
2. Please detive the updating rules for $w^{(1)}$ and $w^{(2)}$ given the classification errors between y and t, where t is the ground truth of the output y.

Question 7 Critical Analyses

1. Please explain why the dual problem formulation is used to solve the SVM machine learning problem.
2. Please explain, in terms of cost functions, constraints and predictions:
    1. what are the differences between SVM classification and logistic regression
    2. what are the differences between ν-SVM regression and least square regression.
3. Please explain why neural network (NN) based machine learning algorithms use logistic activation functions?
4. Please explain:
    1. what are the differences between the logistic activation function and other activation functions (e.g., relu, tanh)
    2. when these activation functions should be used.
5. Please explain why Jacobian and Hessian matrices are useful for machine learning algorithms.
6. Please explain why exponential family distributions are so common in engineering practice.

    Please give some examples which are **NOT** exponential family distributions.
7. Please explain why KL divergence is useful for machine learning? Please provide two examples of using KL divergence in machine learning.
8. Please explain why data augmentation techniques are a kind of regularization skills for NNs.
9. Please explain why Gaussian distributions are preferred over other distributions for many machine learning models?
10. Please explain why Laplace approximation can be used for many cases?
11. What are the fundamental principles for model selection (degree of complexity) in machine learning?

12. How to choose a new data sample (feature) for regression and classification model training, respectively? How to choose it for testing? Please provide some examples.

13. Please explain why the MAP model is usually more preferred than the ML model?

Question 8 Discussions

1. What are the generative and discriminative approaches to machine learning, respectively?

   Can you explain the advantages and disadvantages of these two approaches and provide a detailed example to illustrate your points?

2. How do you analyze the GAN model from the generative and discriminative perspectives?