# ML Test

January 6, 2025

## Research Knowledge Test

- **Please complete all the questions and send your answers in a PDF file generated from Latex code via email within three (3) days of receipt.**
- **Please also include your Latex source code in the email.**
- **Each question is worth 10 points.**

### Statistical learning theory

1. Proof the following conclusion: For a binary classification problem, for all functions in the indicator function set (including the one minimizing the empirical risk), the empirical risk $R_{emp}(w)$ and the expected risk $R(w)$ satisfy the following inequality with at least $1 - \delta$ probability:

$R(w) \leq R_{emp}(w) + \sqrt{\frac{h \ln(2n/h) + \ln(4/\delta)}{n}}$

where $h$ is the VC dimension of the function set, and $n$ is the number of samples.

2. According to VC (Vapnik-Chervonenkis) theory, What factors determine the consistency of empirical risk minimization?

3. As the sample size approaches infinity, what is the relation between the empirical risk $R_{emp}(f)$ and the true risk $R(f)$?

4. Proof your conclusion in Q4

5. Explain the following conclusions and proof them

(1) what is the convergence bound for a single function ?

(2) what is the uniform convergence bound for a finite class of functions ?

(3) what is the uniform convergence bound for both finite and infinite classes of functions?

### Matrix

1. $f = a^T X b$, find $\frac{\partial f}{\partial X}$. Where $a$ is an $m \times 1$ column vector, $X$ is an $m \times n$ matrix, $b$ is an $n \times 1$ column vector, and $f$ is a scalar.

2. $f = a^T \exp(Xb)$, find $\frac{\partial f}{\partial X}$. Where $a$ is an $m \times 1$ column vector, $X$ is an $m \times n$ matrix, $b$ is an $n \times 1$ column vector, exp represents the element-wise exponential, and $f$ is a scalar

3. $f = \mathrm{tr}(Y^T MY)$, $Y = \sigma(WX)$, find $\frac{\partial f}{\partial X}$. Where $W$ is an $\ell \times m$ matrix, $X$ is an $m \times n$ matrix, $Y$ is an $\ell \times n$ matrix, $M$ is an $\ell \times \ell$ symmetric matrix, $\sigma$ represents an element-wise function, and $f$ is a scalar.

4. $l = \|Xw - y\|^2$, find the least squares estimate of $w$, which is equivalent to finding the zero point of $\frac{\partial l}{\partial w}$. Where $y$ is an $m \times 1$ column vector, $X$ is an $m \times n$ matrix, $w$ is an $n \times 1$ column vector, and $l$ is a scalar.

5. Given samples $x_1, \ldots, x_N \sim \mathcal{N}(\mu, \Sigma)$, find the maximum likelihood estimate of the covariance matrix $\Sigma$. The mathematical formula is: $l = \log |\Sigma| + \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^T \Sigma^{-1} (x_i - \bar{x})$, find the zero point of $\frac{\partial l}{\partial \Sigma}$. Here, $x_i$ is an $m \times 1$ column vector, $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ is the sample mean, $\Sigma$ is an $m \times m$ symmetric positive definite matrix, $l$ is a scalar, and log represents the natural logarithm.

6. $l = -\mathbf{y}^T \log \mathrm{softmax}(Wx)$, find $\frac{\partial l}{\partial W}$. Here, $\mathbf{y}$ is an $m \times 1$ column vector with one element equal to 1 and all others equal to 0, $W$ is an $m \times n$ matrix, $x$ is an $n \times 1$ column vector, $l$ is a scalar, and log represents the natural logarithm.

The softmax function is defined as: $\mathrm{softmax}(\mathbf{a}) = \frac{\exp(\mathbf{a})}{\mathbf{1}^T \exp(\mathbf{a})}$,

where $\exp(\mathbf{a})$ represents the element-wise exponential, and $\mathbf{1}$ represents a vector of all ones.

## Analysis

1. Define real numbers using either the Cauchy sequence approach or the Dedekind cut approach.

2. (1) State the Closed Graph Theorem and the Inverse Operator Theorem, and prove that these two theorems are equivalent.

(2) Using the theorem stated in (1), prove that if a linear operator $(A)$ on a Hilbert space $(H)$ satisfies $(\langle \varphi, A\psi \rangle = \langle A\varphi, \psi \rangle)$ for all $(\varphi, \psi \in H)$, then $(A)$ is continuous.

3. Provide the definitions of a normed linear space and a Banach space, and give an example of a Banach space and proof it.

4. (1) Provide the definition of separability. (2) Prove that $(L^\infty[a, b])$ is not a separable space.

5. Prove that the subset $\mathbb{Q}$ of $\mathbb{R}$ is not the intersection of countably many open subsets.

6. Proof the set $[0, 1]$ is uncountable.

7. Let $(X, \mu)$ be a measure space, and let $f_n$ and $f$ be square-integrable functions on it. Prove that $\lim_{n\to+\infty} |f_n - f|2 = 0$ if and only if $\lim n \to +\infty |f_n|_2 = |f|2$ and $(f_n)n \geq 1$ converges to $f$ in measure.

8. Let $(M, \mu)$ be a finite measure space, and define $\delta(f) = \int_M \frac{|f|}{1+|f|}, d\mu, ; f \in (M, \mu)$.

(1) If $g \leq f$, prove that $0 \leq \delta(g) \leq \delta(f)$.

(2) Prove that $\delta(f_1 + f_2) \leq \delta(f_1) + \delta(f_2)$, and that $\delta(f) = 0$ if and only if $f$ is almost zero.

(3) Show that $\delta$ characterizes convergence in measure: $\lim_{i\uparrow\beta} \delta(f_i - f) = 0$ if and only if $(f_i)_{i\uparrow\beta}$ converges to $f$ in measure.

## Deep Learning

1. There is a fully connected neural network with $n$ layer. Let's assume the loss function is the Mean Squared Error. The activate function is Sigmoid function. What are the gradient of $W_l$ and $b_l$, where $W_l$ is the weight of layer $l(\ l < n)$, and $b_l$ is the bias of layer $l(\ l < n)$. Proof your conclusion.

2. If the neural network is CNN and the other conditions remind the same. What are the gradient of $W_l$ and $b_l$ ? Proof your conclusion.