Note on Machine Learning Theory

Zepeng CHEN
The HK PolyU

Date: February 25, 2023

1 Introduction

Definition 1.1 (Supervised Learning)

Definition 1.2 (Unsupervised Learning)

Definition 1.3 (Online Learning)

Definition 1.4 (Mathematical Machine Learning Model)

When we consider machine learning theory rather than classical statistics theory, it is because

- Statistics mostly focuses on asymptotic scenarios, and does not provide non-asymptotic guarantees.
- Statistics requires a well-behaved statistical model, e.g., distribution. However, real data, e.g., text and image, could be more complex.
- Statistics aims to find the entire prob distribution, which could be too costly to compute and analyze.

2 Model-based Statistical Learning

This section covers exponential families, maximum likelihood, method of moments and maximum entropy principle, which overlaps statistics. For more detail, please refer to my note on Advanced Statistics directly.

2.1 Information Theory

Definition 2.1

Given a probability vector $\mathbf{q} = [q_1, \cdots, q_k]$ for a discrete random variable X, the (Shannon) entropy of X is defined as

$$H_{\mathbf{q}}(X) = \sum_{i=1}^{k} q_i log \frac{1}{q_i}.$$

Note on *The entropy value is always non-negative and concave.*

Note on *The entropy measures the uncertainty in a given distribution. Moreover, the entropy is upper-bounded by $logk$ (Jensen's Inequality). Particularly, the upper-bound is achieved by the discrete uniform distribution, i.e., $q_1 = \cdots = q_k = \frac{1}{k}$. This can be proved by solving the entropy maximization problem:*

$$\begin{aligned} \max_{\mathbf{q}\in\mathbb{R}^k} \quad & \sum_{i=1}^{k} q_i log \frac{1}{q_i} \\ s.t. \quad & \sum_{i=1}^{k} q_i = 1, \\ & q_i \geq 0, i = 1, \cdots, k. \end{aligned}$$

Definition 2.2 (Conditional Entropy)

Definition 2.3 (Joint Entropy)

2.2 Maximum Entropy Principle

Definition 2.4 (Empirical Distribution)

Suppose we have n observations such as $x_1, \cdots, x_n$ from an unknown distribution p. The empirical distribution is defined as $\tilde{p} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x = x_i)$.

Definition 2.5 (Maximum Entropy Principle)

Given samples $\mathbf{x}_1, \cdots, \mathbf{x}_n$, we want to find a model in a set of probability distributions

$$M_\phi := \left\{ q \in \mathcal{P}_{\mathcal{X}} : \mathbb{E}_{\hat{\boldsymbol{\theta}}}[\phi(\mathbf{x})] = \frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{x}_i) \right\},$$

conduct the inference and base the decision on the distribution maximizing the Entropy function:

$$\underset{q \in M_\phi}{\text{argmax}} H_q(\mathbf{X}) := \sum_{\mathbf{x} \in \mathcal{X}} q(\mathbf{x}) \log \frac{1}{q(\mathbf{x})}.$$

Note on *This principle chooses the most uncertain model based on the given set M.*

Theorem 2.1

The distribution that maximizes the entropy is an exponential family model with feature function ϕ.

Proof Consider the maximum entropy problem

$$\begin{aligned}
\max_{\mathbf{q}\in\mathbb{R}^{|\mathcal{X}|}} \quad & \sum_{\mathbf{x}\in\mathcal{X}} q_{\mathbf{x}} log \frac{1}{q_{\mathbf{x}}} = -\sum_{\mathbf{x}\in\mathcal{X}} q_{\mathbf{x}} log q_{\mathbf{x}} \\
\text{s.t.} \quad & \sum_{\mathbf{x}\in\mathcal{X}} q_{\mathbf{x}}\phi(\mathbf{x}) = \hat{\boldsymbol{\mu}}, \\
& \sum_{\mathbf{x}\in\mathcal{X}} q_{\mathbf{x}} = 1, \\
& q_{\mathbf{x}} \geq 0, \mathbf{x}\in\mathcal{X},
\end{aligned}$$

as a problem without inequality constraints, i.e.,

$$\begin{aligned}
\max_{\mathbf{q}\in\mathbb{R}^{|\mathcal{X}|}} \quad & -\sum_{\mathbf{x}\in\mathcal{X}} q_{\mathbf{x}} log q_{\mathbf{x}} \\
\text{s.t.} \quad & \sum_{\mathbf{x}\in\mathcal{X}} q_{\mathbf{x}} \begin{bmatrix}\phi(\mathbf{x}) \\ 1\end{bmatrix} = \begin{bmatrix}\hat{\boldsymbol{\mu}} \\ 1\end{bmatrix}.
\end{aligned}$$

Next we consider its Lagrangian problem

$$\mathcal{L}(\mathbf{q},\boldsymbol{\gamma}) = \sum_{\mathbf{x}\in\mathcal{X}} q_{\mathbf{x}} \left(-log q_{\mathbf{x}} - \phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k} - \gamma_{k+1}\right) + \hat{\boldsymbol{\mu}}^\top \boldsymbol{\gamma}_{1:k} + \gamma_{k+1},$$

the stationary KKT condition

$$\nabla_{q_{\mathbf{x}}} \mathcal{L}(\mathbf{q},\boldsymbol{\gamma}) = -log q^*_{\mathbf{x}} - \phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k} - \gamma_{k+1} + 1 = 0$$

leads to

$$q^*_{\mathbf{x}} = exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k} - \gamma_{k+1} + 1\right) \geq 0.$$

Thus, $q^*_{\mathbf{x}}$ is also the optimal solution to the original problem. Moreover,

$$q^*_{\mathbf{x}} \propto exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k}\right)$$

leads to

$$q^*_{\mathbf{x}} = \frac{exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k}\right)}{\sum_{x\in\mathcal{X}} exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k}\right)}$$

due to the constraint that probability $q_{\mathbf{x}}$'s add up to 1. ■

Note on *Suppose $\Omega = \{0,1\}$ and $\phi(x) = x$, then the maximum entropy principle leads to a binomial distribution. Suppose $\Omega = (-\infty,\infty)$ and $\phi(x) = [x, x^2]^\top$, then the maximum entropy principle leads to a Gaussian distribution.*

2.3 Maximum Relative Entropy Principle (meila_stat_2012)

Definition 2.6

Based on the logic of maximum entropy principle, suppose we also have a prior distribution q_0.

2.4 Minimum KL-Divergence

Definition 2.7 (Kullback-Leibler Divergence)

Let X be a random variable with possible outcomes $\mathcal{X}$ and let P and Q be two probability distributions on X. The KL-Divergence of P from Q is defined as:

$$KL[P||Q] = \sum_{x\in\mathcal{X}} p(x)log_b\frac{p(x)}{q(x)},$$

$$KL[P||Q] = \int_{\mathcal{X}} p(x)log_b\frac{p(x)}{q(x)}dx.$$

Note on *KL-Divergence is asymmetric in* (p, q).

Lemma 2.1

KL-Divergence is always non-negative. Moreover, $KL(p||q) = 0$ *iff* $p = q$.

Proof By Gibbs' Inequality. ■

Lemma 2.2

KL-Divergence is convex in the pair of probability distributions (p, q)*, i.e.,*

$$KL[\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2] \leq \lambda KL[p_1||q_1] + (1-\lambda)KL[p_2||q_2],$$

where (p_1, q_1) *and* (p_2, q_2) *are two pairs of probability distributions and* $0 \leq \lambda \leq 1$. *Moreover, KL-divergence attains the minimum* 0 *if* $p = q$.

Proof Firstly, we show that KL-Divergence is bi-convex of p, q, i.e., it is a convex function of p for a fixed q and vice versa,

$$KL[\lambda p_1 + (1-\lambda)p_2||q_1] \leq \lambda KL[p_1||q_1] + (1-\lambda)KL[p_2||q_1],$$

$$KL[p_1||\lambda q_1 + (1-\lambda)q_2] \leq \lambda KL[p_1||q_1] + (1-\lambda)KL[p_1||q_2]$$

on the basis of $xlogx$'s convexity. Next we can prove the convexity from bi-convexity. ■

2.5 Connections

2.5.1 Maximum Entropy and MLE

Theorem 2.2 (Maximum Entropy v.s. MLE)

The maximum entropy problem over M_ϕ is the dual problem to the MLE for the exponential family with feature function ϕ.

Proof Suppose we replace the Lagrangian problem with $q^*_{\mathbf{x}}$, the dual of the maximum entropy problem can be written as

$$\min_{\gamma} log\left(\sum_{x\in\mathcal{X}} exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k}\right)\right) + \hat{\boldsymbol{\mu}}^\top \boldsymbol{\gamma}_{1:k}.$$

■

2.5.2 Minimum KL-Divergence and MLE

Theorem 2.3

MLE is minimizing KL-Divergence, i.e.,

$$\arg\min_\theta KL(p||q) = \arg\max_\theta p(x|\theta).$$

Proof

$$\begin{aligned}\arg\min_\theta KL(p||q) = \arg\min_\theta \mathbb{E}_{x\sim p}\left[log\frac{p(x)}{q(x)}\right] \\ \iff \arg\min_\theta \mathbb{E}_{x\sim p}\left[-logq(x)\right] \\ \iff \arg\max_\theta \mathbb{E}_{x\sim p}\left[logq(x)\right]\end{aligned}$$

■

3 Model-free Machine Learning

Compared to model-based learning, here the underlying distribution is unknown. We introduce the theory of supervised learning in this section.

Definition 3.1

A standard goal in supervised learning is to minimize the averaged prediction loss $\ell : \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}^{\geq 0}$ where $\ell(\hat{y}, y)$ measures the loss suffered under prediction $\hat{y}$ for actual label y.

Note on Example of Loss Function

- *Squared-error loss:* $\ell_2(\hat{y}, y) = (\hat{y} - y)^2$,
- *0-1 loss:* $\ell_{0/1}(\hat{y}, y) = I(\hat{y} \neq y)$.

Definition 3.2 (Supervised Learning)

Given ℓ, the supervised learning goal is to find a prediction function $f \in \mathcal{F}$ to minimize the expected loss under distribution $P_{X,Y}$, i.e., population risk:

$$\min_{f \in \mathcal{F}} \mathbb{E}_{P_{X,Y}}[\ell(f(X), Y)].$$

Note on *This problem cannot be solven, since we do not know $P_{X,Y}$.*

Definition 3.3 (Empirical Risk Minimization)

3.1 Uniform Convergence Bounds

3.2 VC Dimension

3.3 Rademacher Complexity

3.4 Covering Numbers

4 Theory of Representation

4.1 Kernel Functions and Methods

4.2 Approximation in Deep Learning

5 Theory of Convergence