Machine Learning

Notes taken by Runqiu Ye
Carnegie Mellon University

Spring 2025

Contents

1 Probability and Statistical Inference

1.1 Probability

Definition (Types of convergence). Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and X be another random variable. Let F_n be the CDF of X_n for each $n \in \mathbb{N}$ and F be the CDF of X.

1. X_n converges to X *in probability* and write $X_n \xrightarrow{\text{P}} X$ if for arbitrary $\varepsilon > 0$,
$$\mathbb{P}\left[|X_n - X| > \varepsilon\right] \to 0$$
as $n \to \infty$.

2. X_n converges to X *in distribution* and write $X_n \rightsquigarrow X$ if
$$\lim_{n\to\infty} F_n(t) = F(t)$$
for all t where F is continuous.

3. X_n converges to X in L^p if
$$\mathbb{E}\left[|X_n - X|^p\right] \to 0$$
as $n \to \infty$. In particular, say X_n converges to X in *quadratic mean* and write $X_n \xrightarrow{\text{qm}} X$ if X_n converges to X in L^2.

4. X_n converges to X *almost surely* and write $X_n \xrightarrow{\text{as}} X$ if
$$\mathbb{P}\left[\lim_{n\to\infty} X_n = X\right] = 1.$$

Theorem. The following implication holds:

1. If X_n converges to X almost surely, then X_n converges to X in probability.
2. If X_n converges to X in L^p, then X_n converges to X in probability.

Proof. 1. If X_n converges to X almost surely, the set of points $O = \{\omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\}$ has measure zero. Now fix $\varepsilon > 0$ and consider the sequence of sets
$$A_n = \bigcup_{m=n}^{\infty} \left\{|X_m - X| > \varepsilon\right\}.$$
Note that $A_n \supset A_{n+1}$ for each $n \in \mathbb{N}$ and let $A_\infty = \bigcap_{n=1}^{\infty} A_n$. Now show $\mathbb{P}[A_\infty] = 0$. If $\omega \notin O$, then $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ and thus $|X_n(\omega) - X(\omega)| < \varepsilon$ for some $n \in \mathbb{N}$. Therefore, $\omega \notin A_\infty$. It follows that $A_\infty \subset O$ and $\mathbb{P}[A_\infty] = 0$.

By monotone continuity, we have $\lim_{n\to\infty} \mathbb{P}[A_n] = \mathbb{P}[A_\infty]$. It follows that
$$\mathbb{P}\left[|X_n - X| > \varepsilon\right] \le \mathbb{P}\left[A_n\right] \to 0$$
as $n \to \infty$. This completes the proof.

2. From Chebyshev's inequality, we have
$$\mathbb{P}\left[|X - X_n| > \varepsilon\right] \le \frac{1}{\varepsilon^p}\mathbb{E}[|X - X_n|^p].$$
The claim follows directly.

□

Theorem (Central Limit Theorem). Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Let $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then
$$Z_n = \frac{S_n - \mu}{\sqrt{\operatorname{Var} S_n}} = \frac{\sqrt{n}\,(S_n - \mu)}{\sigma} \rightsquigarrow Z,$$
where $Z \sim N(0,1)$. In other words,
$$\lim_{n\to\infty} \mathbb{P}[Z_n < z] = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx.$$
Also write $Z_n \approx N(0,1)$.

1.2 Statistical Inference

Definition. Let $X_1, \dots, X_n$ be n i.i.d. data points from some distribution F. A point estimator $\widehat{\theta}_n$ of a parameter θ is some function of $X_1, \dots, X_n$:

$$\widehat{\theta}_n = g(X_1, \dots, X_n).$$

The bias of an estimator is defined as

$$\text{bias}(\widehat{\theta}_n) = \mathbb{E}_\theta[\widehat{\theta}_n] - \theta.$$

The mean squared error is defined as

$$\text{MSE} = \mathbb{E}_\theta(\widehat{\theta}_n - \theta)^2.$$

Definition. A point estimator $\widehat{\theta}_n$ of a parameter θ is *consistent* if $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$.

Theorem. The MSE can be written as

$$\text{MSE} = \text{bias}^2(\widehat{\theta}_n) - \text{Var}_\theta(\widehat{\theta}_n).$$

Definition. A $1-\alpha$ interval for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \dots, X_n)$ and $b = b(X_1, \dots, X_n)$ are functions of data such that

$$\mathbb{P}_\theta[\theta \in C_n] \geq 1 - \alpha \text{ for all } \theta \in \Theta.$$

In other word, (a, b) traps θ with probablity $1-\alpha$.

Warning! In the above definition, C_n is random and θ is fixed.

2 Supervised Learning

2.1 Logistic Regression

Logistic regression is used for classfication problems. Logistic regression takes in input feature $x \in \mathbb{R}^n$, and output a prediction $y \in \{0, 1\}$. The hypotheses function $h_\theta(x)$ is chosen as

$$h_\theta(x) = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}},$$

where

$$g(z) = \frac{1}{1 + e^{-z}}$$

is the sigmoid function.

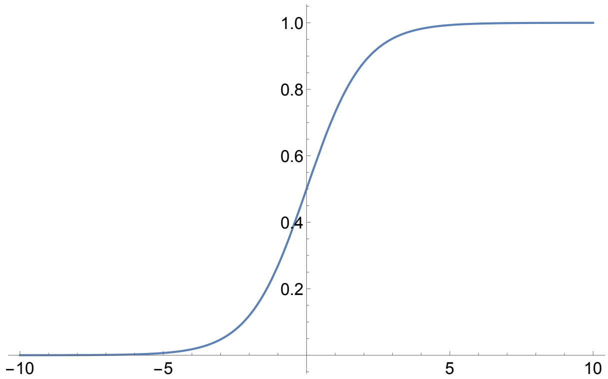

Figure 1: A plot of the sigmoid function $\sigma(z)$.

A plot of the sigmoid function is shown in Figure 1. The range of the sigmoid function is bounded in $[0, 1]$. In particular, $\sigma(z) \to 1$ when $z \to \infty$ and $\sigma(z) \to 0$ as $z \to -\infty$. A useful property about the sigmoid function is its derivative. It is easy to verify that

$$\sigma'(z) = \frac{e^{-z}}{(1 + e^{-z})^2} = \sigma(z)(1 - \sigma(z)).$$

To fit the parameter θ to dataset, assume that

$$\begin{aligned} p(y = 1 \mid x; \theta) &= h_\theta(x), \\ p(y = 0 \mid x; \theta) &= 1 - h_\theta(x). \end{aligned}$$

Note that

$$p(y \mid x; \theta) = h_\theta(x)^y (1 - h_\theta(x))^{1-y}.$$

Assuming n independent training examples, the likelihood function

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta) \\ &= \prod_{i=1}^{n} h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}. \end{aligned}$$

It is easier to maximize the log-likelihood:

$$\ell(\theta) = \sum_{i=1}^{n} y^{(i)} h_\theta(x^{(i)}) + (1 - y^{(i)})(1 - h_\theta(x^{(i)})).$$

This is called the logisitic loss or the binary cross-entropy.