Frequentist statistics

A short review

Roulin Wang

School of Management
University of Science and Technology of China

August 05, 2020

Table of Contents

Table of Contents

Introduction

Frequentist statistics (sometimes called classical statistics or orthodox statistics) is an approach to statistics. The polar opposite is Bayesian statistics.

- "Parameters" are fixed but unknown and data are random

抛硬币的试验

试验者	掷硬币的次数	正面出现的次数	频率
蒲丰	4040	2048	.5069
皮尔逊	12000	6019	.5016
皮尔逊	24000	12012	.5005

Introduction

Frequentist statistics (sometimes called classical statistics or orthodox statistics) is an approach to statistics. The polar opposite is Bayesian statistics.

- Probability is a measure of frequency of repeated events. That is, the probability of an event A is the limit

$$P(A) = \lim_{n \to \infty} \frac{\#\text{times observed A}}{\#\text{times we looked}}$$

抛硬币的试验

试验者	掷硬币的次数	正面出现的次数	频率
蒲丰	4040	2048	.5069
皮尔逊	12000	6019	.5016
皮尔逊	24000	12012	.5005

framework

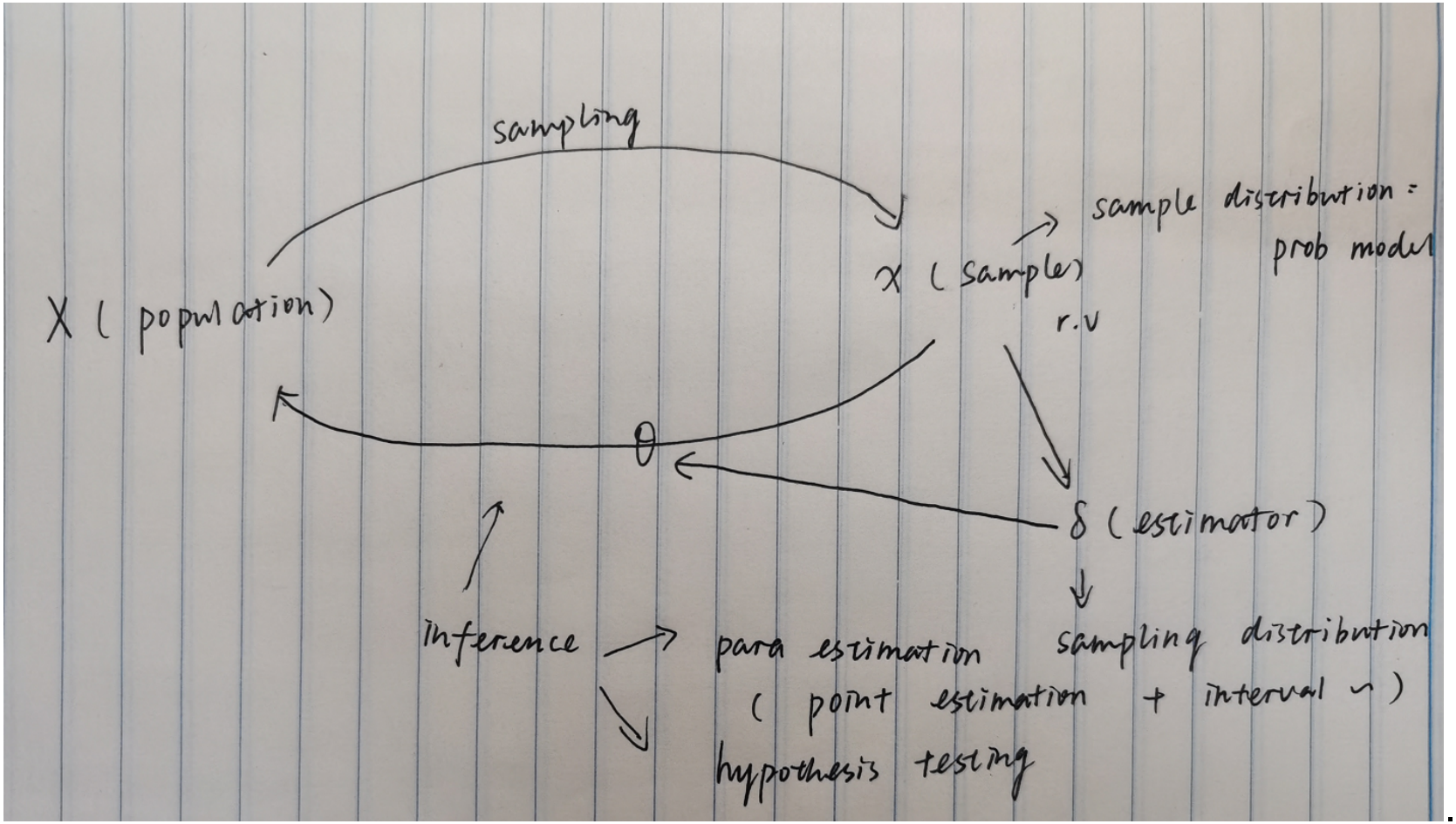

sampling
X (population)
X (sample)
r.v
sample distribution =
prob model
θ
δ (estimator)
inference
para estimation
(point estimation + interval)
hypothesis testing
sampling distribution

Table of Contents

Sampling distribution of an estimator

- In frequentist statistics, $\hat{\theta} = \delta(D)$, D is the sample data.
- The uncertainty in the parameter estimate can be measured by computing the sampling distribution of the estimator.

Sampling distribution with a normal population

When the population distribution is normal, sampling distributions of many important statistics have been given, and these are mostly closely related to the following three distributions:

- Chi-Square (χ^2) distribution : Let $Z_1, Z_2, \cdots, Z_\nu$ be $N(0,1)$ r.v.'s and let $X = Z_1^2 + Z_2^2 + \cdots + Z_\nu^2$. Then the pdf of X can be shown to be:

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} \quad \text{for } x \geq 0.$$

  This is the χ^2 distribution with ν degrees of freedom.

- t distribution : Let Z be $N(0,1)$, $Y \sim \chi^2_n$. X and Y are independent. Let $t = \frac{Z}{\sqrt{Y/n}}$. The pdf of T is

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma(\nu/2)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} \quad -\infty < t < \infty.$$

  This is the t distribution with n df.

Sampling distribution with a normal population

- F distribution : Consider r.v. X and Y are independent, and $X \sim \chi^2_m$, $Y \sim \chi^2_n$. Let $W = \frac{X}{m} / \frac{Y}{n}$, then W has a F-distribution, $W \sim F_{m,n}$. The pdf of W is:

$$f(w) = \frac{\Gamma\left((m+n)/2\right)}{\Gamma\left(m/2\right)\Gamma\left(n/2\right)} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1 + \frac{m}{n} w\right)^{-(m+n)/2}$$

$$\text{for } w \geq 0.$$

Sampling mean distribution CLT

- Let $X_1, X_2, \cdots, X_n$ be a random sample drawn from distribution with a finite mean μ and variance σ^2. As $n \to \infty$, the distribution of:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

"converges" to the distribution $N(0, 1)$.

- "converges" means "converges in distribution"

$$\lim_{n\to\infty} P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) = \Phi(z) \text{ for all } z.$$

- CLT characterizes large samples from any distribution. As long as you have a lot of independent samples (from any distribution), then the distribution of the sample mean is approximately normal.

Large sample theory for the MLE

Under certain conditions, as the sample size tends to infinity, the sampling distribution of the MLE becomes Gaussian.

- The center of the Gaussian will be the MLE $\hat{\theta}$.
- Intuitively, the variance of the estimator will be (inversely) related to the amount of curvature of the likelihood surface at its peak.

Large sample theory for the MLE

Formalize this intuition!

- score function:

$$\mathbf{s}(\hat{\boldsymbol{\theta}}) \triangleq \nabla \log p(\mathcal{D} \mid \boldsymbol{\theta})\Big|_{\hat{\boldsymbol{\theta}}}.$$

- observed information matrix :

$$\mathbf{J}(\hat{\boldsymbol{\theta}}(\mathcal{D})) \triangleq -\nabla \mathbf{s}(\hat{\boldsymbol{\theta}}) = -\nabla^2_{\boldsymbol{\theta}} \log p(\mathcal{D} \mid \boldsymbol{\theta})\big|_{\hat{\boldsymbol{\theta}}}.$$

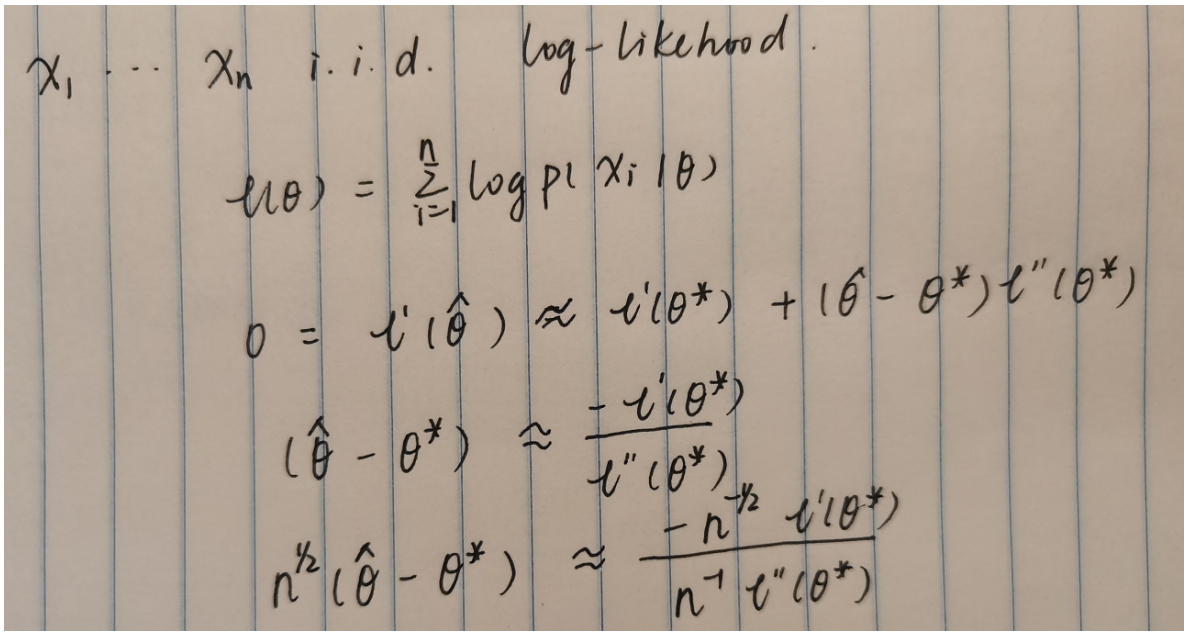
- Fisher information matrix :

$$\mathbf{I}_N\left(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}^*\right) \triangleq \mathbb{E}_{\boldsymbol{\theta}^*}[\mathbf{J}(\hat{\boldsymbol{\theta}} \mid \mathcal{D})].$$

Large sample theory for the MLE

Important theorem

Under smoothness conditions, the probability distribution of $\sqrt{nI(\theta^*)}\left(\hat{\theta}-\theta^*\right)$ tends to a standard normal distribution.



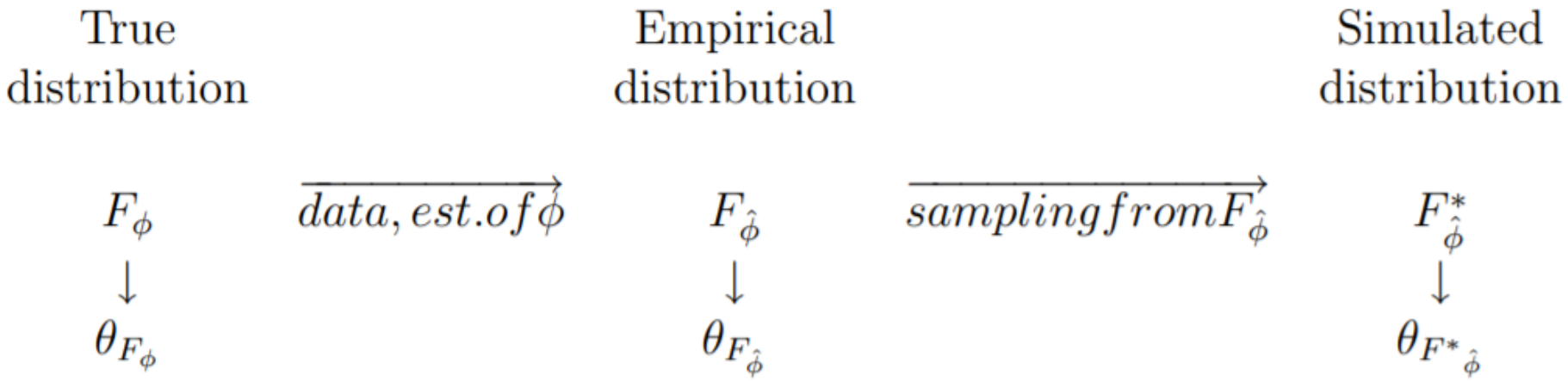
Remark: "smoothness conditions" means that the order of differentiation and integration of p can be interchanged.

Bootstrap

The bootstrap is a simple Monte Carlo technique to approximate the sampling distribution.

- Nonparametric bootstrap : The population distribution is unknown, use the re-sampling method to (re)sample from the sample for statistical inference.
- Parametric bootstrap : The population distribution is known, using Monte Carlo's method to sample from the distribution for statistical inference.

Parametric bootstrap



Example: Empirical exponential distribution
Let$X_i \sim Exp(\lambda)$, for all $i = 1, \ldots, n$, where λ is unknown. The c.d.f. F and the density function f of an exponential r.v. are following:

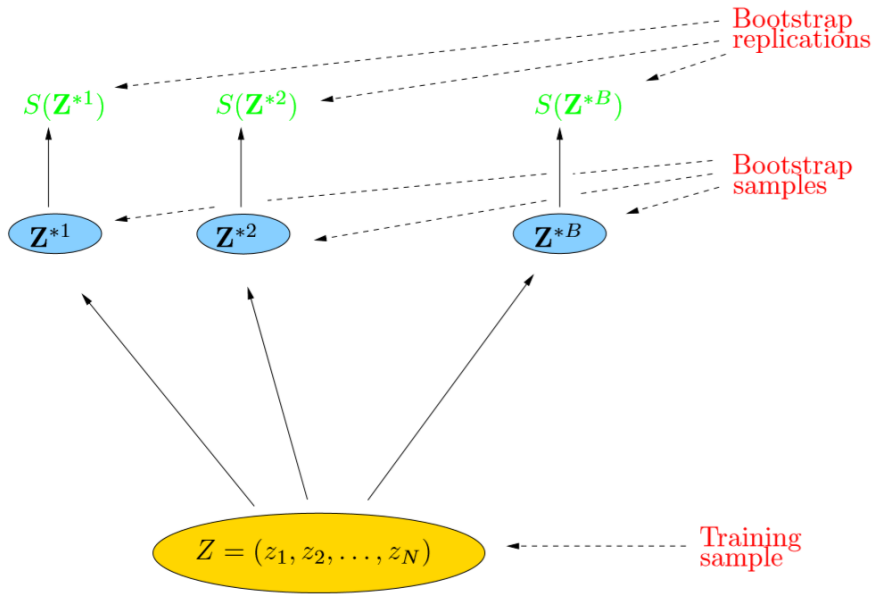
$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-\lambda x} & \text{if } x > 0 \end{cases}$$

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \lambda e^{-\lambda x} & \text{if } x > 0 \end{cases}$$

We know that $E(X_i) = \frac{1}{\lambda}$ and $\bar{X}$ is an unbiased estimator of $\frac{1}{\lambda}$. Hence, $\hat{\lambda} = \frac{1}{\bar{x}}$ may be used to get the empirical distribution, i.e., $F_{\hat{\lambda}} = F_{\frac{1}{\bar{x}}}$. So, we may generate samples from $Exp(\frac{1}{\bar{x}})$.

Nonparametric Bootstrap

- training set: Z
- Sampling times (with replacement): B
- quantity (computed from the data): $S(Z)$



Bootstrap

From the bootstrap sampling, we can estimate any aspect of the distribution of $S(Z)$.

- mean: $\bar{S}^* = \sum_b S\left(\mathbf{Z}^{*b}\right)/B$.
- variance: $\widehat{\mathrm{Var}}[S(\mathbf{Z})] = \frac{1}{B-1}\sum_{b=1}^{B}\left(S\left(\mathbf{Z}^{*b}\right) - \bar{S}^*\right)^2$.
- prediction error: $\widehat{\mathrm{Err}}_{\mathrm{boot}} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{*b}\left(x_i\right)\right)$.

where $\hat{f}^{*b}\left(x_i\right)$ is the predicted value at x_i, from the model fitted to the bth bootstrap dataset.

Bootstrap Confidence Intervals

Method of constructing asymptotic confidence interval of target parameters in bootstrap.

- The Standard Normal Bootstrap Confidence Interval
- The Percentile Bootstrap Confidence Interval
- The Basic Bootstrap Confidence Interval
- The Bootstrap t interval
- Better Bootstrap Confidence Intervals

Jackknife

- The jackknife (or leave one out) method, invented by Quenouille (1949), is an alternative resampling method to the bootstrap.
- The method is based upon sequentially deleting one observation from the dataset, recomputing the estimator, here, $\hat{\theta}_{(i)}$, n times. That is, there are exactly n jackknife estimates obtained in a sample of size n.
- Like the bootstrap, the jackknife method provides a relatively easy way to estimate the precision of an estimator, θ.
- The jackknife is generally less computationally intensive than the bootstrap.

Jackknife Algorithm

- For a dataset with n observations, compute n estimates by sequentally omitting each observation from the dataset and estimating $\hat{\theta}$ on the remaining $n-1$ observations.
- Using the n jackknife estimates, $\hat{\theta}_{(1)}, \hat{\theta}_{(2)}, \ldots, \hat{\theta}_{(n)}$, we estimate : the bias of the estimator as :

$$\widehat{\text{bias}}_{jack} = (n-1)(\overline{\hat{\theta}_{(\cdot)}} - \hat{\theta}).$$

The standard error of the estimator as

$$\widehat{se}_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \overline{\hat{\theta}}_{(\cdot)}\right)^2},$$

where $\overline{\hat{\theta}}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$.

boostrap and jackknife

Bootstrap Summary

Advantages

- All purpose computer intensive method useful for statistical inference.
- Bootstrap estimates of precision do not require knowledge of the theoretical form of an estimator's standard error, no matter how complicated it is.

Disadvantages

- Typically not useful for correlated (dependent) data.
- Missing data, censoring, data with outliers are also problematic
- Often used incorrectly

Jackknife Summary

Advantages

- Useful method for estimating and compensating for bias in an estimator.
- Like the bootstrap, the methodology does not require knowledge of the theoretical form of an estimator's standard error.
- Is generally less computationally intensive compared to the bootstrap method.

Disadvantages

- The jackknife method is more conservative than the bootstrap method, that is, its estimated standard error tends to be slightly larger.
- Performs poorly when the the estimator is not sufficiently smooth, i.e., a non-smooth statistic for which the jackknife performs poorly is the median.

Table of Contents

Desirable properties of estimators

- Unbiased estimation
- Minimum variance estimation
- The bias-variance trade-off
- Consistent estimation

Unbiased estimators

The bias of an estimator is defined as

$$\text{bias}(\hat{\theta}(\cdot)) = \mathbb{E}_{p(\mathcal{D}|\theta_*)}\left[\hat{\theta}(\mathcal{D}) - \theta_*\right],$$

where $\theta_\star$ is the true parameter value. If the bias is zero, the estimator is called unbiased 0

- Suppose $\mathcal{D} = \{x_1, \ldots, x_N\}$ obeys Gaussian distribution $N(\mu, \sigma^2)$.

$$bias(\bar{x}) = bias(x_1) = 0.$$

Minimum variance estimators

The variance of the $\bar{x}$ and x_1 mentioned above are :

$$variance(\bar{x}) = \frac{1}{N}\sigma^2, \quad variance(x_1) = \sigma^2.$$

the variance of an estimator is also important.
how long can the variance go?

Cramer-Rao inequality

Let $X_1, \cdots, X_n \sim p(X|\theta_0)$ and $\hat{\theta} = \hat{\theta}(x_1, \cdots, x_n)$ be an unbiased estimator of θ_0. Then, under various smoothness assumptions on $p(X|\theta_0)$, we have

$$\text{var}[\hat{\theta}] \geq \frac{1}{nI(\theta_0)},$$

where $I(\theta_0)$ is the Fisher information matrix.

UMVUE

- Uniformly Minimum-Variance Unbiased Estimator (UMVUE): an unbiased estimator that has lower variance than any other unbiased estimator for all possible values of the parameter .

The bias-variance trade-off

Examples

Example: estimating a Gaussian mean, assume the data is sampled from $x_i \sim N(\theta^\star = 1, \sigma^2)$

- MLE estimator: $\bar{x}$.
- MAP estimator: $\tilde{x} \triangleq \frac{N}{N+\kappa_0}\bar{x} + \frac{\kappa_0}{N+\kappa_0}\theta_0 = w\bar{x} + (1-w)\theta_0$,

$$\mathbb{E}[\tilde{x}] - \theta^* = w\theta_0 + (1-w)\theta_0 - \theta^* = (1-w)(\theta_0 - \theta^*),$$

$$\text{var}[\tilde{x}] = w^2 \frac{\sigma^2}{N}.$$

The bias-variance trade-off

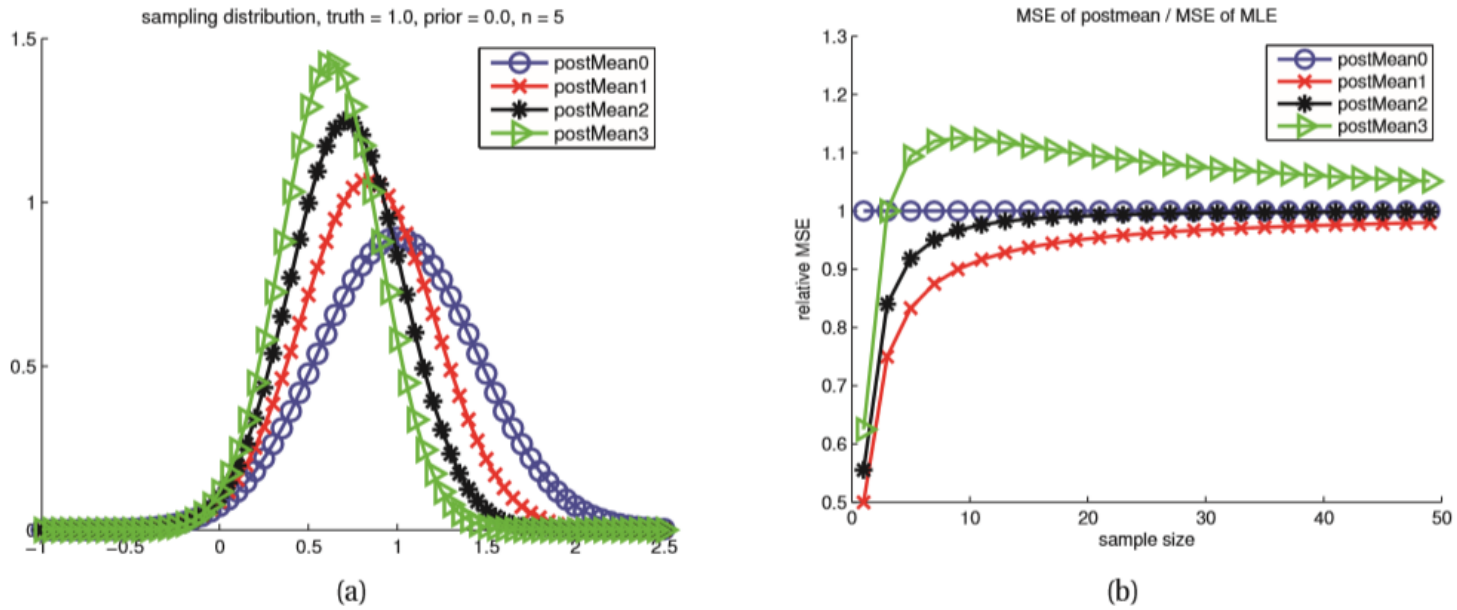


Figure 6.4 Left: Sampling distribution of the MAP estimate with different prior strengths κ_0. (The MLE corresponds to $\kappa_0 = 0$.) Right: MSE relative to that of the MLE versus sample size. Based on Figure 5.6 of (Hoff 2009). Figure generated by `samplingDistGaussShrinkage`.

The bias-variance trade-off

use quadratic loss, the corresponding risk is the MSE.

$$\begin{aligned}
\mathbb{E}\left[(\hat{\theta}-\theta^*)^2\right] &= \mathbb{E}\left[\left[(\hat{\theta}-\bar{\theta})+(\bar{\theta}-\theta^*)\right]^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\bar{\theta}\right)^2\right]+2(\bar{\theta}-\theta^*)\mathbb{E}\left[\hat{\theta}-\bar{\theta}\right]+(\bar{\theta}-\theta^*)^2 \\
&= \mathbb{E}\left[\left(\hat{\theta}-\bar{\theta}\right)^2\right]+(\bar{\theta}-\theta^*)^2 \\
&= \text{var}\left[\hat{\theta}\right]+\text{bias}^2(\hat{\theta})
\end{aligned}$$

In words,

MSE = variance + bias^2

This is called the bias-variance tradeoff.

Consistent estimators

An estimator is said to be consistent if it eventually recovers the true parameters that generated the data as the sample size goes to infinity, i.e.,$\hat{\theta}(\mathcal{D}) \to \theta^*$ as $|\mathcal{D}| \to \infty$.

Table of Contents

Confidence Intervals

- Interval estimation is an alternative to the variety of techniques we have examined.
- Given data x, we replace the point estimate $\hat{\theta}(x)$ for the parameter θ by a statistic that is subset $\hat{C}(x)$ of the parameter space.
- the random set $\hat{C}(x)$ is chosen to have a prescribed high probability, γ, of containing the true parameter value θ.

$$P_\theta(\theta \in \hat{C}(x)) = \gamma.$$

In this case, the set $\hat{C}(x)$ is called a γ-level confidence interval.

Methods of derivation

(Neyman) methods for constructing confidence intervals.

- Constructing confidence interval by point estimator.
  Often take the form

$$\hat{C}(\mathbf{x}) = (\hat{\theta}(\mathbf{x}) - m(\mathbf{x}), \hat{\theta}(\mathbf{x}) + m(\mathbf{x})) = \hat{\theta}(\mathbf{x}) \pm m(\mathbf{x}),$$

  where $\hat{\theta}(x)$ is a point estimate, and $m(x)$ is the margin of error.

Methods of derivation

- Hypothesis testing. Consider the hypothesis problem

$$H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0$$

, there is a test at level α of the hypothesis. Denote the acceptance region of the test by $A(\theta_0)$. Then the set

$$C(\mathbf{X}) = \{\theta : \mathbf{X} \in A(\theta)\}$$

is a $100(1-\alpha)$ confidence region for θ.

- The Duality of Confidence Intervals and Hypothesis Tests (Mathematical Statistics and Data Analysis).

Examples

Example 16.1 (1-sample z interval). *If $X_1.X_2.\ldots.X_n$ are normal random variables with unknown mean μ but known variance σ_0^2. Then,*

$$Z = \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}}$$

Examples

Let $X_1, \ldots, X_n$ be a random sample from a normal distribution having unknown mean μ and known variance σ^2. We consider testing the following hypotheses:

$$H_0: \mu = \mu_0$$
$$H_A: \mu \neq \mu_0$$

Consider a test at a specific level α that rejects for $|\overline{X} - \mu_0| > x_0$, where x_0 is determined so that $P(|\overline{X} - \mu_0| > x_0) = \alpha$ if H_0 is true: $x_0 = \sigma_{\overline{X}} z(\alpha/2)$. Here the standard deviation of $\overline{X}$ is denoted by $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. The test thus accepts when

$$|\overline{X} - \mu_0| < \sigma_{\overline{X}} z(\alpha/2)$$

or

$$-\sigma_{\overline{X}} z(\alpha/2) < \overline{X} - \mu_0 < \sigma_{\overline{X}} z(\alpha/2)$$

or

$$\overline{X} - \sigma_{\overline{X}} z(\alpha/2) < \mu_0 < \overline{X} + \sigma_{\overline{X}} z(\alpha/2)$$

A $100(1-\alpha)\%$ confidence interval for μ_0 is

$$\left[\overline{X} - \sigma_{\overline{X}} z(\alpha/2), \overline{X} + \sigma_{\overline{X}} z(\alpha/2)\right]$$

Table of Contents

Frequentist decision theory

- Decision space : $\mathcal{A}, \theta^\star \in \mathcal{A}$.
- Observed data : $X_1, \cdots, X_n \sim (\mathcal{X}^n, P_{\theta^\star})$.
- Target : $\theta^\star$ (density estimation).
- Decision : $\delta : \mathcal{X}^n \to \mathcal{A}$.
- Loss function :
$$L : \mathcal{A} \times \mathcal{A} \to \mathbb{R}^+.$$
- Risk function :
$$R(\theta^\star, \delta) = \mathbb{E}_{\theta^\star} L(\delta, \theta^\star).$$

Frequentist decision theory

In frequentist or classical decision theory, there is a loss function and a likelihood.

- In the Bayesian approach, the Bayesian posterior expected loss

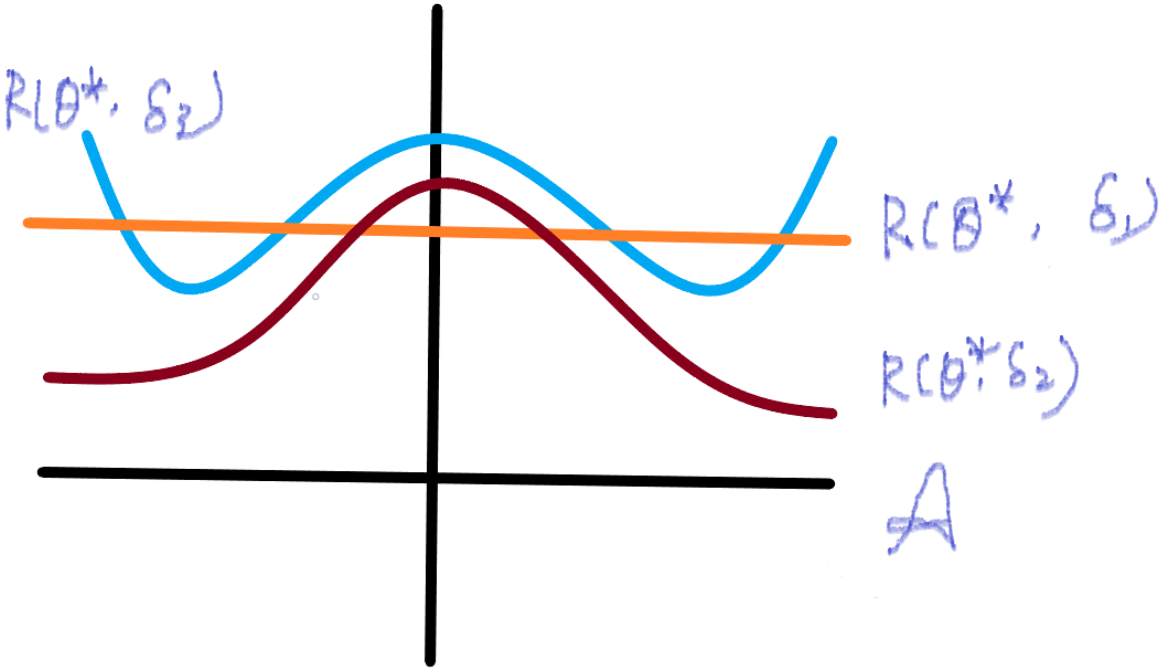
$$\rho(a \mid \mathcal{D}, \pi) \triangleq \mathbb{E}_{p(\theta|\mathcal{D},\pi)}[L(\theta, a)] = \int_{\Theta} L(\theta, \mathbf{a}) p(\theta \mid \mathcal{D}, \pi) d\theta.$$

- In the frequentist approach, choose decision procedure $\delta : \mathcal{X}^n \to \mathcal{A}$, define its expected loss or risk as follows

$$R\left(\theta^{\star}, \delta\right) \triangleq \mathbb{E}_{p\left(\tilde{\mathcal{D}}|\theta^{\star}\right)}\left[L\left(\theta^{\star}, \delta(\tilde{\mathcal{D}})\right)\right] = \int L\left(\theta^{\star}, \delta(\tilde{\mathcal{D}})\right) p\left(\tilde{\mathcal{D}} \mid \theta^{\star}\right) d\tilde{\mathcal{D}}.$$

Not only is the frequentist definition unnatural, it cannot even be computed.

Frequentist decision theory



Question

Is δ_1 better than δ_2?
Is δ_3 admissible?

Bayes risk

Convert $R(\theta^\star, \delta)$ into a single measure of quality $R(\delta)$.
Average case : If we have some measure of $\mathcal{A} : \Pi$,

$$\int R(\theta^\star, \delta)\Pi(d\theta^\star)$$

could be used in comparison, where Π is called the prior of $\mathcal{A}$.
The optimal δ_B:

$$\operatorname*{argmin}_{\delta} \int R(\theta^\star, \delta)\Pi(d\theta^\star)$$

exists under some mild conditions, which is the functionals of posterior distribution : $\Pi(\theta|X_{1:n})$. For this reason, optimal decision is often called Bayes procedure

Bayes risk

Theorem 6.3.1

A Bayes estimator can be obtained by minimizing the posterior expected loss for each x.

$$
\begin{aligned}
R_B(\delta) &= \int R(\theta^*, \delta)\, p(\theta^*)\, d\theta^* \\
&= \int \left[\int L(\theta^*, \delta(x))\, p(x \mid \theta^*) \right] p(\theta^*)\, d\theta^* \\
&= \sum_x \left[\int L(\theta^*, \delta(x))\, p(x \mid \theta^*)\, p(\theta^*)\, d\theta^* \right] \\
&= \sum_x \left[\int L(\theta^*, \delta(x))\, p(\theta^* \mid x)\, d\theta^* \right] p(x) \\
&= \sum_x p(\delta(x) \mid x)\, p(x)
\end{aligned}
$$

Minimax risk

Define the maximum risk of an estimator as

$$R_{\max}(\delta) \triangleq \max_{\boldsymbol{\theta}^*} R\left(\boldsymbol{\theta}^*, \delta\right).$$

A minimax rule is one which minimizes the maximum risk:

$$\delta_{MM} \triangleq \operatorname*{argmin}_{\delta} R_{\max}(\delta).$$

Admissible estimators

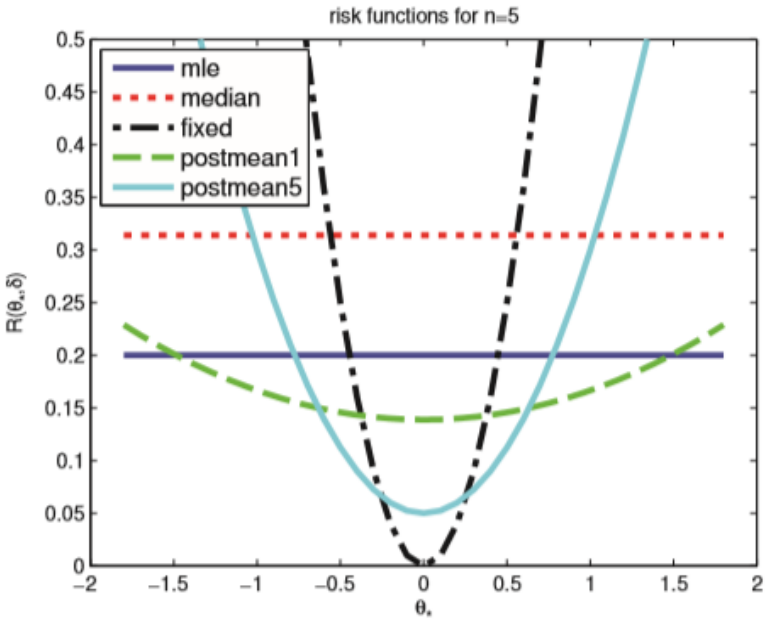
- Def : if $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all $\theta \in \Theta$, then we say that δ_1 dominates δ_2. An estimator is said to be admissible if it is not strictly dominated by any other estimator.

Let us give an example, based on (Bernardo and Smith 1994). Consider the problem of estimating the mean of a Gaussian. We assume the data is sampled from $x_i \sim \mathcal{N}(\theta^*, \sigma^2 = 1)$ and use quadratic loss, $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. The corresponding risk function is the MSE. Some possible decision rules or estimators $\hat{\theta}(\mathbf{x}) = \delta(\mathbf{x})$ are as follows:

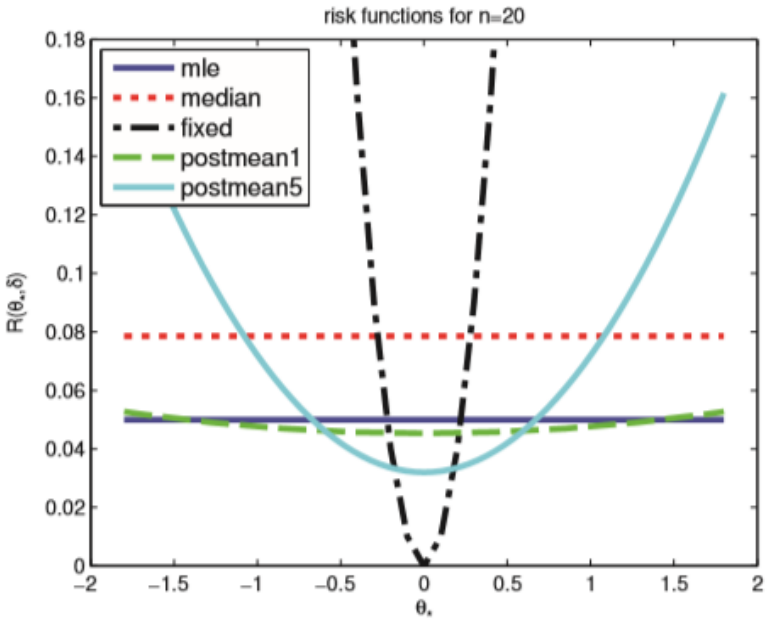
- $\delta_1(\mathbf{x}) = \overline{x}$, the sample mean
- $\delta_2(\mathbf{x}) = \tilde{\mathbf{x}}$, the sample median
- $\delta_3(\mathbf{x}) = \theta_0$, a fixed value
- $\delta_\kappa(\mathbf{x})$, the posterior mean under a $\mathcal{N}(\theta|\theta_0, \sigma^2/\kappa)$ prior:

$$\delta_\kappa(\mathbf{x}) = \frac{N}{N+\kappa}\overline{x} + \frac{\kappa}{N+\kappa}\theta_0 = w\overline{x} + (1-w)\theta_0 \tag{6.20}$$

Admissible estimators



(a)



(b)

Stein's paradox

- Let $X_1, X_2, \cdots, X_n$ be independent p-dimensional normal random variables, such that $X_i \sim N(\boldsymbol{\theta}, I_p)$. Consider estimating the mean $\boldsymbol{\theta}$ under squared error loss $L(\boldsymbol{\theta}, \mathbf{d}) = \|\mathbf{d} - \boldsymbol{\theta}\|^2 = (\mathbf{d} - \boldsymbol{\theta})^\top (\mathbf{d} - \boldsymbol{\theta})$.
- Is the naive estimator $\hat{\boldsymbol{\theta}} = \bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$ admissible?

Stein's paradox

The Answer is YES when $p = 1, 2$.
But when $p > 2$, James-Stein show that the obvious estimate $\bar{\mathbf{X}}$ is dominated by

$$\hat{\boldsymbol{\theta}}_{\mathrm{JS}} = \left(1 - \frac{p-2}{\|\overline{\mathbf{X}}\|^2}\right) \overline{\mathbf{X}}.$$

This is known as Stein's paradox.

Admissibility is not enough

Theorem 6.3.3.

Let $X \sim N(\theta, 1)$, and consider estimating θ under squared loss. Let $\delta_1(x) = x_0$, a constant independent of the data. This is an admissible estimator.

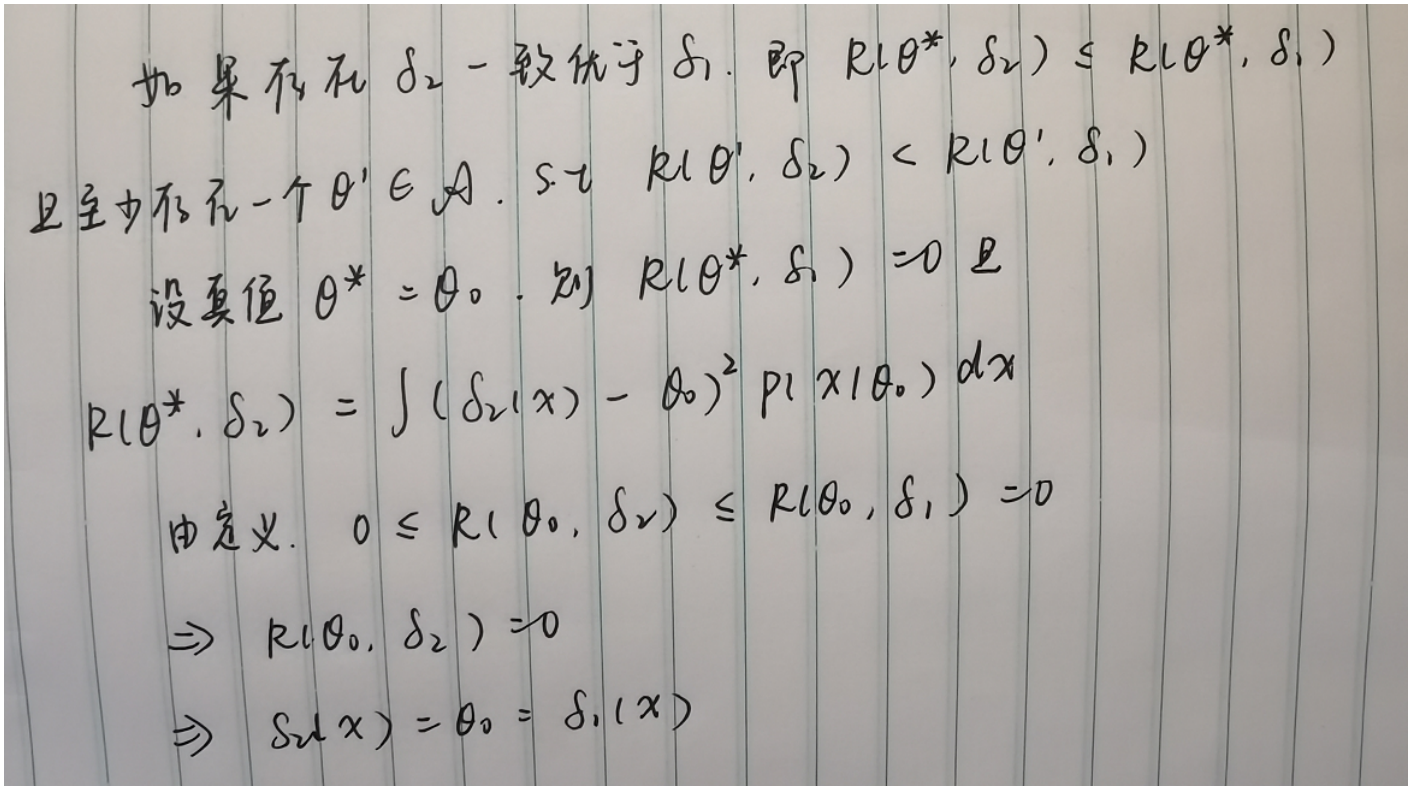


Table of Contents

Empirical risk minimization

Frequentist decision theory suffers from the fundamental problem that one cannot actually compute the risk function.

Loss functions of the form $L(y, \delta(x))$, the frequentist risk becomes

$$R\left(p_*, \delta\right) \triangleq \mathbb{E}_{(\mathbf{x},y)\sim p_*}[L(y, \delta(\mathbf{x})] = \sum_{(\mathbf{x},y)} L(y, \delta(\mathbf{x}))p_*(\mathbf{x}, y).$$

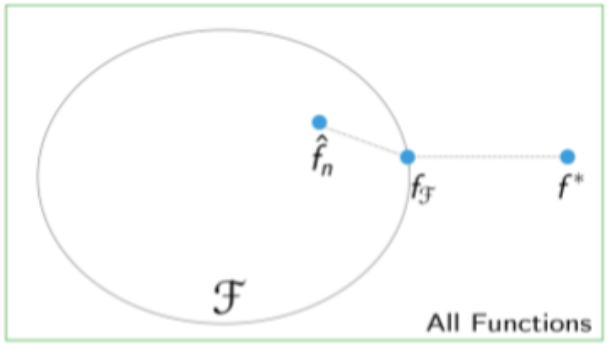
where,

$$p_*(\mathbf{x}, y) \approx p_{\text{emp}}(\mathbf{x}, y) \triangleq \frac{1}{N}\sum_{i=1}^{N} \delta_{\mathbf{x}_i}(\mathbf{x})\delta_{y_i}(y).$$

Then define the empirical risk as follows:

$$R_{emp}(\mathcal{D}, \mathcal{D}) \triangleq R\left(p_{\text{emp}}, \delta\right) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \delta\left(\mathbf{x}_i\right)\right).$$

ERM in machine learning



$$f^* = \arg\min_f \mathbb{E}\ell(f(X), Y)$$

$$f_{\mathcal{F}} = \arg\min_{f\in\mathcal{F}} \mathbb{E}\ell(f(X), Y))$$

$$\hat{f}_n = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$$

- **Approximation Error** (of $\mathcal{F}$) $= R(f_{\mathcal{F}}) - R(f^*)$
- **Estimation error** (of $\hat{f}_n$ in $\mathcal{F}$) $= R(\hat{f}_n) - R(f_{\mathcal{F}})$

- The excess risk of the ERM $\hat{f}_n$ can be decomposed:

$$\begin{aligned} \textbf{Excess Risk}(\hat{f}_n) &= R(\hat{f}_n) - R(f^*) \\ &= \underbrace{R(\hat{f}_n) - R(f_{\mathcal{F}})}_{\text{estimation error}} + \underbrace{R(f_{\mathcal{F}}) - R(f^*)}_{\text{approximation error}}. \end{aligned}$$

Regularized risk minimization

regularized risk minimization (RRM):

$$R'(\mathcal{D}, \delta) = R_{emp}(\mathcal{D}, \delta) + \lambda C(\delta),$$

where $C(\delta)$ measures the complexity of the prediction function $\delta(x)$ and λ controls the strength of the complexity penalty.

The two key issues : how do we measure complexity, and how do we pick λ.

Structural risk minimization

structural risk minimization principle:

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}} \hat{R}\left(\hat{\delta}_{\lambda}\right),$$

where $\hat{R}(\delta)$ is an estimate of the risk.

There are two widely used estimates: cross validation and theoretical upper bounds on the risk.

Estimating the risk using cross validation

Let F be a learning algorithm or fitting function that takes a dataset and a model index m .

- returns a parameter vector: $\hat{\boldsymbol{\theta}}_m = \mathcal{F}(\mathcal{D}, m)$.
- returns a prediction: $\hat{y} = \mathcal{P}(\mathbf{x}, \hat{\boldsymbol{\theta}}) = f(\mathbf{x}, \hat{\boldsymbol{\theta}})$.
- the cross-validation estimate of prediction error is:

$$\mathrm{CV}(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, \hat{f}^{-\kappa(i)}\left(x_i\right)\right).$$

- K-fold CV estimate of the risk of f_m is:

$$R(m, \mathcal{D}, K) \triangleq \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} L\left(y_i, \mathcal{P}\left(\mathbf{x}_i, \mathcal{F}\left(\mathcal{D}_{-k}, m\right)\right)\right),$$

where $f_m(\mathbf{x}, \mathcal{D}) = \mathcal{P}(\mathbf{x}, \mathcal{F}(\mathcal{D}, m))$.

Upper bounding the risk using statistical learning theory

Question

Are there analytic approximations or bounds to the generalization error?

Consider the case where the hypothesis space is finite, with size $\dim(H) = |H|$, $R(p_*, h)$ is the risk function for any data distribution p_* and hypothesis $h \in \mathcal{H}$, $R_{emp}(\mathcal{D}, h)$ is the empirical risk . Then we have the following.

Theorem 6.5.1. *For any data distribution p_*, and any dataset $\mathcal{D}$ of size N drawn from p_*, the probability that our estimate of the error rate will be more than ϵ wrong, in the worst case, is upper bounded as follows:*

$$P\left(\max_{h\in\mathcal{H}} |R_{emp}(\mathcal{D}, h) - R(p_*, h)| > \epsilon\right) \leq 2\dim(\mathcal{H})e^{-2N\epsilon^2} \tag{6.66}$$