Chapter 6
Frequentist statistics

Attempts have been made to devise approaches to statistical inference that avoid treating parameters like random variables, and which thus avoid the use of priors and Bayes rule. Such approaches are known as **frequentist statistics**, **classical statistics** or **orthodox statistics**. Instead of being based on the posterior distribution, they are based on the concept of a sampling distribution.

6.1 Sampling distribution of an estimator

In frequentist statistics, a parameter estimate $\hat{\theta}$ is computed by applying an **estimator** δ to some data $\mathcal{D}$, so $\hat{\theta} = \delta(\mathcal{D})$. The parameter is viewed as fixed and the data as random, which is the exact opposite of the Bayesian approach. The uncertainty in the parameter estimate can be measured by computing the **sampling distribution** of the estimator. To understand this

6.1.1 Bootstrap

We might think of the bootstrap distribution as a poor mans Bayes posterior, see (Hastie et al. 2001, p235) for details.

6.1.2 Large sample theory for the MLE *

6.2 Frequentist decision theory

In frequentist or classical decision theory, there is a loss function and a likelihood, but there is no prior and hence no posterior or posterior expected loss. Thus there is no automatic way of deriving an optimal estimator, unlike the Bayesian case. Instead, in the frequentist approach, we are free to choose any estimator or decision procedure $f : \mathcal{X} \to \mathcal{Y}$ we want.

Having chosen an estimator, we define its expected loss or **risk** as follows:

$$\begin{aligned} R_{\exp}(\theta, f) &\triangleq \mathbb{E}_{p(\tilde{\mathcal{D}}|\theta^*)}[L(\theta^*, f(\tilde{\mathcal{D}}))] \\ &= \int L(\theta^*, f(\tilde{\mathcal{D}})) p(\tilde{\mathcal{D}}|\theta^*) \mathrm{d}\tilde{\mathcal{D}} \end{aligned} \tag{6.1}$$

where $\tilde{\mathcal{D}}$ is data sampled from natures distribution, which is represented by parameter θ^*. In other words, the expectation is wrt the sampling distribution of the estimator. Compare this to the Bayesian posterior expected loss:

$$\rho(f|\mathcal{D},) \tag{6.2}$$

6.3 Desirable properties of estimators

6.4 Empirical risk minimization

6.4.1 Regularized risk minimization

6.4.2 Structural risk minimization

6.4.3 Estimating the risk using cross validation

6.4.4 Upper bounding the risk using statistical learning theory *

6.4.5 Surrogate loss functions

log-loss

$$L_{\mathrm{nll}}(y, \eta) = -\log p(y|\boldsymbol{x}, \boldsymbol{w}) = \log(1 + e^{-y\eta}) \tag{6.3}$$

6.5 Pathologies of frequentist statistics *