Statistical Learning: Problem Set 1

Instituto Superior Técnico

April 20, 2020

Problem 1 - Frequentist Decision Theory

Consider the problem of estimating the parameters μ and σ^2 of a Gaussian random variable $\mathrm{X} \sim \mathcal{N}(\mu, \sigma^2)$ given the observation of n random variables $\mathrm{X}_1, ..., \mathrm{X}_n$ that are i.i.d with X. We start by employing the frequentist approach with loss function $L\big(s, \delta(\mathrm{X}_1, ..., \mathrm{X}_n)\big) = \big(s - \delta(\mathrm{X}_1, ..., \mathrm{X}_n)\big)^2$.

1) Estimation of μ.

a) Use Chebyshev's inequality to prove that $\delta_1\big(\mathrm{X}_1, ..., \mathrm{X}_n\big) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{X}_i$ is a consistent estimator of μ. Derive a sufficient condition for an unbiased estimator to be consistent.

b) Let $\delta_2\big(\mathrm{X}_1, ..., \mathrm{X}_n\big) = 4$ be another estimator of μ. Compute its bias and variance, and prove that δ_2 is an admissible estimator. *Suggestion: start by computing the risk of δ_2 when the true parameter μ is 4.*

2) Estimation of σ^2 when μ is known.

Suppose that μ is a known constant. Derive the maximum likelihood estimator for σ^2 and compute its expectation. Comment the result.

3) Estimation of σ^2 when μ is unknown.

a) Let $\bar{\mathrm{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{X}_i$. Show that $\delta_3\big(\mathrm{X}_1, ..., \mathrm{X}_n\big) = \frac{1}{n}\sum_{i=1}^{n}\big(\mathrm{X}_i - \bar{\mathrm{X}}\big)^2$ is a biased estimator of σ^2 and conclude that $\delta_4\big(\mathrm{X}_1, ..., \mathrm{X}_n\big) = \frac{1}{n-1}\sum_{i=1}^{n}\big(\mathrm{X}_i - \bar{\mathrm{X}}\big)^2$ is an unbiased estimator of σ^2.

b) Compute the Cramer-Rao bound for estimating σ^2 and the variance of $\delta_4\big(\mathrm{X}_1, ..., \mathrm{X}_n\big)$. What do you conclude?

Problem 2 - Bayesian Decision Theory

1) Reject option in classifiers.

(Source: (Duda et al. 2001, Q2.13).)

In many classification problems one has the option either of assigning $\mathbf{x}$ to class j or, if you are too uncertain, of choosing the **reject option**. If the cost for rejects is less than the cost of falsely classifying the object, it may be the optimal action. Let α_i mean you choose action i, for $i = 1 : C + 1$, where C is the number of classes and $C + 1$ is the reject action. Let $Y = j$ be the true (but unknown) **state of nature**. Define the loss function as follows

$$\lambda(\alpha_i|Y = j) = \begin{cases} 0 & \text{if } i = j \text{ and } i, j \in \{1, \ldots, C\} \\ \lambda_r & \text{if } i = C + 1 \\ \lambda_s & \text{otherwise} \end{cases} \tag{5.122}$$

In otherwords, you incur 0 loss if you correctly classify, you incur λ_r loss (cost) if you choose the reject option, and you incur λ_s loss (cost) if you make a substitution error (misclassification).

Decision	true label y	
$\hat{y}$	0	1
predict 0	0	10
predict 1	10	0
reject	3	3

a. Show that the minimum risk is obtained if we decide $Y = j$ if $p(Y = j|\mathbf{x}) \geq p(Y = k|\mathbf{x})$ for all k (i.e., j is the most probable class) *and* if $p(Y = j|\mathbf{x}) \geq 1 - \frac{\lambda_r}{\lambda_s}$; otherwise we decide to reject.

b. Describe qualitatively what happens as λ_r/λ_s is increased from 0 to 1 (i.e., the relative cost of rejection increases).

2) MPM estimator for variance of a Gaussian r.v..

Complete the skipped calculations of slide 36 (Lecture 2) and conclude that the MPM (*maximizer of posterior marginals*) estimator for the variance of a Gaussian random variable is unbiased. Start by computing the posterior marginal of the variance at point σ^2, that is, $f_{\Sigma^2|\mathbf{X}}(\sigma^2|x)$.

Problem 3 - Frequentist vs Bayesian

1) The Beta-Binomial model.

Suppose you wish to estimate the probability p that a coin lands on heads. For that purpose, consider the random variable X that is equal to one if the coin lands on heads, and is equal to zero otherwise, *i.e.*, $\mathbb{P}(\mathrm{X}=1) = 1-\mathbb{P}(\mathrm{X}=0) = p$. Suppose the coin is flipped n times, and the variables $\mathrm{X}_1, ..., \mathrm{X}_n$ corresponding to each flip are observed. Naturally, the variables $\mathrm{X}_1, ..., \mathrm{X}_n$ are i.i.d. with X.

a) Derive the *maximum likelihood* (ML) estimator for p. Prove that it is indeed, a maximum of the likelihood function, not just a stationary point.

b) If we wish to express some degree of certainty about p using a Bayesian approach, we must do so by selecting a prior for p, denoted f_P. Naturally, since $p \in [0,1]$, we must choose a prior with adequate support, *i.e.*, $\{p \in \mathbb{R} : f_P(p) > 0\} \subseteq [0,1]$. The Beta distribution is an excellent candidate, since it is a conjugate prior for both the Bernoulli and Binomial distributions.

i. Suppose you wish that the prior conveys the information that the expectation of p is 0.6. What is the family of Beta distributions that satisfy this condition?

ii. Assume a Beta prior for p, that is, $p \sim \mathrm{Beta}(\alpha, \beta)$. Ignoring the normalizing constant, derive the posterior density $f_{P|\mathrm{X}_1,...,\mathrm{X}_n}(p|x_1, ..., x_n)$ and the MAP estimate of p. What is the MAP estimate of p if the coin is flipped 4 times, and every observation was heads?

c) Optional. The so-called Jeffrey's prior for p is $p \sim \mathrm{Beta}(1/2, 1/2)$. This prior has the property of being invariant under reparametrizations (*i.e.*, invertible continuous transformations of the parameter). Show that this is indeed the case if, instead of p, the Bernoulli is parameterized with the log-odds ratio $\eta = \log(p/(1-p))$.

2) MMSE vs MAP.

Consider the problem of estimating the mean of a Gaussian random variable $\mathrm{X} \sim \mathcal{N}(m, \sigma^2)$ with known variance, given a set of observations $\mathrm{X}_1, ..., \mathrm{X}_n$ i.i.d. with X. Since m is unknown, it is modeled as a random variable M with some prior f_M. Let $\mathrm{Z}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathrm{Z}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be two Gaussian random variables, with known $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$. Let $p \in (0,1)$ and suppose M satisfies

$$M = \begin{cases} \mathrm{Z}_1, & \text{with probability } p \\ \mathrm{Z}_2, & \text{with probability } 1-p \end{cases}.$$

a) Use the law of total probability to deduce the density of the prior, f_M.

b) Compute the MMSE and MAP estimators of m. Which of those has a convenient analytical expression?

Problem 4 - Beyond i.i.d. Observations

You are watching an ongoing marathon where every competitor has a unique number on their shirt, ranging from 1 to N. You don't know how many competitors there are, so you don't know N. You observe four people running, with the numbers $n_1, ..., n_4$, which are randomly sampled from the N competitors with equal probability. Derive the maximum likelihood estimate for N. Generalize to the case where you observe K samples: $n_1, ..., n_K$.