ECE 271A: Homework #3

Due on November 13, 2023 at 11:59pm

Professor Vasconcelos

Ray Tsai

A16848188

Problem 1

In this problem we will consider the issue of linear regression and the connections between maximum likelihood and least squares solutions. Consider a problem where we have two random variables Z and X, such that

$$z = f(x, \theta) + \epsilon \tag{1}$$

where f is a polynomial with parameter vector θ

$$f(x, \theta) = \sum_{k=0}^{K} \theta_k x^k$$

and ϵ a Gaussian random variable of zero mean and variance σ^2. Our goal is to estimate the best estimate of the function given i.i.d. sample $\mathcal{D} = \{(\mathcal{D}_x, \mathcal{D}_z)\} = \{(x_1, z_1), \dots, (x_n, z_n)\}$.

Part A

Formulate the problem as one of least squares, i.e define $z = (z_1, \dots, z_n)^T$,

$$\Phi = \begin{bmatrix} 1 & \dots & x_1^K \\ \vdots & \ddots & \vdots \\ 1 & \dots & x_n^K \end{bmatrix}$$

and find the value of θ that minimizes

$$\|z - \Phi\theta\|^2.$$

Solution

We attempt to find θ, such that it gives a closest solution to

$$\Phi\theta = z.$$

By performing least squares, we get

$$\theta = (\Phi^T \Phi)^{-1} \Phi^T z.$$

Part B

Formulate the problem as one of ML estimation, i.e. write down the likelihood function $P_{Z|X}(z|x;\theta)$, and compute the ML estimate, i.e. the value of θ that maximizes $P_{Z|X}(\mathcal{D}_z|\mathcal{D}_x;\theta)$. Show that this is equivalent to part A.

Solution

Suppose that X is known. Then, $P_{Z|X}(z|x;\theta)$ become a Gaussian distribution with mean $f(x, \theta)$ and variance σ^2, namely

$$P_{Z|X}(z|x;\theta) = G(x, f(x, \theta), \sigma^2).$$

Problem 1 continued on next page. . .

Given sampple $\mathcal{D}$, we take the natural log of $P_{Z|X}(\mathcal{D}_z|\mathcal{D}_x;\theta)$ and get

$$\begin{aligned}\theta^* &= \arg\max_\theta \sum_{i=1}^n -\frac{(z_i - f(x_i,\theta))^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\\ &= \arg\min_\theta \sum_{i=1}^n (z_i - f(x_i,\theta))^2\\ &= \arg\min_\theta \|z - \Phi\theta\|^2,\end{aligned}$$

and what we're looking for is obviously identical to the question in part A.

Part C

(The advantage of the statistical formulation is that makes the assumptions explicit. We will now challenge some of these.) Assume that instead of a fixed variance σ^2 we now have a variance that depends on the sample point, i.e.

$$z_i = f(x_i,\theta) + \epsilon_i,$$

where $\epsilon_i \sim N(0,\sigma_i^2)$. This means that our sample is independent but no longer identically distributed. It also means that we have different degrees of confidence in the different measurements (z_i, x_i). Redo part B under these conditions.

Solution

Instead of looking at individual data points, we view $\mathcal{D}_z$ and $\mathcal{D}_x$ as random vectors. Thus, the probablity distribution becomes a Gaussian distribution with mean $\Phi\theta$ and variance $\Sigma = \operatorname{diag}(\sigma_1^2,\ldots,\sigma_n^2)$, namely

$$P_{Z|X}(z|\mathcal{D}_x;\theta) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp -\frac{1}{2}(z-\Phi\theta)^T\Sigma^{-1}(z-\Phi\theta).$$

We again do the log trick and get

$$\begin{aligned}\theta^* &= \arg\max_\theta \ln\left(\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\right) - \frac{1}{2}(z-\Phi\theta)^T\Sigma^{-1}(z-\Phi\theta)\\ &= \arg\min_\theta\,(z-\Phi\theta)^T\Sigma^{-1}(z-\Phi\theta).\end{aligned}$$

Let $g(\theta) = (z-\Phi\theta)^T\Sigma^{-1}(z-\Phi\theta)$. We take the gradient of g with respect to θ and get

$$\nabla_\theta g = -2\Phi^T\Sigma^{-1}(z-\Phi\theta) = 0.$$

Thus, we get a critical point $\theta^* = (\Phi^T\Sigma^{-1}\Phi)^{-1}\Phi^T\Sigma^{-1}z$. We take the Hessian of g and get that

$$\nabla_\theta^2 g = 2\Phi^T\Sigma^{-1}\Phi = 2(S\Phi)^T(S\Phi),$$

where $S = diag(\sigma_1^{-1},\ldots,\sigma_n^{-1})$. Since $\nabla_\theta^2 g$ can be decomposed into a product of a matrix and its transpose, it is positive definite, and so θ^* is the minimum point.

Problem 1 continued on next page. . .

Part D

Consider the weighted least squares problem where the goal is to minimize

$$(z - \Phi\theta)^T W (z - \Phi\theta),$$

where W is a symmetrix matrix. Compute the optimal θ in this situation. What is the equivalent maximum likelihood problem? Rewrite the model (1), making explicit all the assumptions that lead to the new problem. What is the statistical interpretation of W?

Solution

By part C, we know the least square solution to this problem is

$$\theta^* = (\Phi^T W \Phi)^{-1} \Phi^T W z.$$

We can thus assume that $W = \Sigma^{-1}$ is the inverse of the covariance matrix, such that the random noise vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \sim N(0, \Sigma)$. Thus, (1) can be rewritten into

$$z = \Phi\theta + \epsilon,$$

where $z = (z_1, \ldots, z_n)^T$.

Part E

The L_2 norm is known to be prone to large estimation error if there are outliers in the training sample. These are training examples (z_i, x_i) for which, due to measurement errors or other extraneous causes, $|z_i - \sum_k \theta_i x_i^k|$ is much larger than for the remaining examples (the *inliers*). In fact, it is known that a single outlier can completely derail the least squares solution, an highly undesirable behavior. It is also well known that other norms lead to much more robust estimators. One of such distance metrics is the L_1-norm

$$L_1 = \sum_i \left| z_i - \sum_k \theta_k x_i^k \right|.$$

In the maximum likelihood framework, which is the statistical assumption that leads to the L_1 norm? Once again, rewrite the model (1), making explicit all the assumptions that lead to the new problem. Can you justify why this alternative formulation is more robust? In particular, provide a justification for i) why the L_1 norm is more robust to outliers, and ii) the associated statistical model (1) copes better with them.

Solution