UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: Probability and Random Processes

Discussion 12
Fall 2020

1. MMSE for Jointly Gaussian Random Variables

Provide justification for each of the following steps (1 - 5) to prove that the LLSE is equal to the MMSE estimator for jointly Gaussian random variables X and Y.
Let $g(X) = L[Y \mid X]$.

$$
\begin{aligned}
& E[(Y - g(X))X] = 0 && (1)\\
\implies\ & \operatorname{cov}(Y - g(X), X) = 0 && (2)\\
\implies\ & Y - g(X) \text{ is independent of } X && (3)\\
\implies\ & E[(Y - g(X))f(X)] = 0\ \forall f(\cdot) && (4)\\
\implies\ & g(X) = E[Y \mid X] && (5)
\end{aligned}
$$

2. Orthogonal LLSE

(a) Consider zero-mean random variables X, Y, Z such that Y, Z are orthogonal. Show that $L[X \mid Y, Z] = L[X \mid Y] + L[X \mid Z]$.

(b) Explain why for any zero-mean random variables X, Y, Z it holds that:

$$L[X \mid Y, Z] = L[X \mid Y] + L\big[X \;\big|\; Z - L[Z \mid Y]\big]$$

3. Fun with Linear Regression

Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown linear function, i.e. it is of the form $f(x) = x^\top w = x_1 w_1 + \cdots + x_d w_d$, where $w \in \mathbb{R}^d$ is the unknown parameter of the linear function. We pick n points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$, and we observe $y^{(1)}, \ldots, y^{(n)} \in \mathbb{R}$ that are generated according to the model

$$y^{(i)} = f\big(x^{(i)}\big) + \epsilon_i, \text{ for } i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables.

Let us first estimate w when we have **no prior** information about it.

(a) Compute the likelihood of the parameter w given the data $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$

$$\mathcal{L}(w \mid \{(x^{(i)}, y^{(i)})\}_{i=1}^n) := \prod_{i=1}^n p\big(y^{(i)} \mid x^{(i)}; w\big).$$

(b) Explicitly define a matrix $X \in \mathbb{R}^{n \times d}$ and a vector $y \in \mathbb{R}^n$ such that the optimal points of the problem

$$\min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2,$$

correspond to the maximizers of the likelihood.

Now assume a zero-mean **Gaussian prior** for each $w_i,\ i = 1, \ldots, d$. In particular assume that $w_1, \ldots, w_d$ are i.i.d. $\mathcal{N}(0, \tau^2)$, and they are also independent of the data.

(c) Compute, up to a normalization constant, the posterior distribution of w given the data $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$.

(d) Explicitly define a matrix $X \in \mathbb{R}^{n \times d}$, a vector $y \in \mathbb{R}^n$ and a positive scalar $\lambda \in \mathbb{R}$ such that the optimal point of the problem

$$\min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2 + \lambda \|w\|_2^2,$$

correspond to the maximizer of the posterior distribution of w.