EE660 Homework 1 (Fall 2025)

Assigned: 9/2/2025, Due: 9/16/2025

Instructions: You may collaborate with others on this problem set, but each student must independently write their own solutions. We highly encourage you to use LaTeX to typeset your solutions. We will accept handwritten assignments; however if the solution to a problem is too illegible for the grader to read, then they may use their discretion and consider the problem incomplete. Solutions are due by 11:59pm Pacific Time on the due date, and are only to be submitted on Brightspace. Do not email the course staff with your assignment.

Template: The link `https://www.overleaf.com/read/hjgknqqhryqy` contains a basic LaTeX template that you may use. Note, however, that you are not required to use this template.

GPT Policy: Please review the GPT usage policy from the course syllabus: `https://stephentu.github.io/pdfs/EE660_Fa2025_Syllabus.pdf`.

Note: In the problem set below, Lecture Notes refer to the course lecture notes. The latest version is here: `https://stephentu.github.io/pdfs/EE660_Lecture_Notes.pdf`.

1. Problem 1

Exercise 1.2 in the Lecture Notes.

2. Problem 2

Exercise 1.6 in the Lecture Notes.

3. Problem 3

Exercise 1.7 in the Lecture Notes.

4. Problem 4

In lecture, we saw how applying duality theory to the primal SVM problem yields a dual SVM problem which only depends on inner products. In this problem, we will show that this phenomenon does not only happen for SVMs, but also for least-squares classification problems.

(a) Let $S_n = \{(x_i, y_i)\}_{i=1}^n$ be a training dataset with $x_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$. Suppose we learn a predictor of the form $f_w(x) = \mathrm{sgn}(\langle w, x \rangle)$, with $w \in \mathbb{R}^d$, by minimizing the following least-squares loss:[1]

$$\hat{w}_n = \arg\min_{w \in \mathbb{R}^d} \hat{L}_P(w) := \frac{1}{2n} \sum_{i=1}^n (\langle x_i, w \rangle - y_i)^2 = \frac{1}{2n} \|Xw - Y\|^2, \tag{P}$$

where $X \in \mathbb{R}^{n \times d}$ is the covariate matrix with the i-th row of X equal to $x_i \in \mathbb{R}^d$, and $Y = (y_1, \ldots, y_n) \in \{\pm 1\}^n$ (make sure to convince yourself that the two expressions above are indeed equal). Now consider the following alternative optimization procedure:[2]

$$\hat{\alpha}_n = \arg\min_{\alpha \in \mathbb{R}^n} \hat{L}_D(\alpha) := \frac{1}{2n} \|XX^\mathsf{T}\alpha - Y\|^2. \tag{D}$$

Show that (P) and (D) both obtain the same objective value, i.e., show that:

$$\hat{L}_P(\hat{w}_n) = \hat{L}_D(\hat{\alpha}_n).$$

(b) Use part (a) to design a *kernel* least-squares classification algorithm. Specifically, given as input (a) training data S_n and (b) a positive definite kernel function $k(x, y)$, describe an algorithm to learn a predictor of the form $f_\alpha(x) = \mathrm{sgn}(\sum_{i=1}^n k(x, x_i)\alpha_i)$ by solving a least-squares regression problem for $\alpha \in \mathbb{R}^n$.

Hint 1: Do not simply copy the algorithm for kernel SVMs from the lecture/lecture notes, that is *not* what we are looking for.

Hint 2: When $k(x, y) = \langle x, y \rangle$ is the linear kernel, your proposed least-squares regression for computing the weights α should be identical to (D).

5. Problem 5

Exercise 1.16 in the Lecture Notes.

Extra Hint: In lecture, we saw that if k_1, k_2 are valid kernels, then their sum $k_1 + k_2$ is as well. Immediately, this implies that for any finite N, the function $\bar{k}_N(x, y) := \sum_{i=1}^N k_i(x, y)$ is a valid kernel whenever all the k_i's are valid. Less obvious but still true is that, given a countably infinite number of kernels $\{k_i\}_{i=1}^\infty$, as long as the sum $\sum_{i=1}^\infty k_i(x, y)$ converges pointwise for every (x, y), then $\bar{k}(x, y) := \sum_{i=1}^\infty k_i(x, y)$ is also a valid kernel; this is a consequence of the fact that the positive semidefinite cone is a closed set. You may use this fact without proof.

[1] When the minimizer to (P) is not unique, we will take $\hat{w}_n$ to be one of the minimizers. How we select such a minimizer is a topic we will come back to in the future, but for now we will skip over this detail and assume we have some strategy in place.

[2] The same comment regarding the non-uniqueness of optimizers for (P) also applies to optimizers for (D).