HW3

Kevin Chang

October 13, 2025

1

Recall that for a vector $w \in \mathbb{R}^d$, $\mathcal{H}_w := \{z :< w, z >= 0\}$. Let $S = \{(x_i, y_i)\}$ be a set of linearly separable data in $\mathbb{R}^d$ (i.e., $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$). Define the set $\mathcal{M}_S$ to be the set of all vectors which separate the data with large dot product:

$$\mathcal{M}_S = \{w : y_i < w, x_i >\geq 1 \text{for} i = 1, ..., n\}.$$

- Let w^* denote the element of $\mathcal{M}_S$ with smallest norm. Show that for any other w that separates the data
$$\min dist_{1\leq i\leq n}(x_i, \mathcal{H}_w) \leq \min_{1\leq i\leq n} dist(x_i, \mathcal{H}_{w^*}).$$

  Recall that for any nonzero vector $w \in \mathbb{R}^d$, the distance from a point x to the hyperplane $\mathcal{H}_w := \{z : \langle w, z\rangle = 0\}$ is given by

$$\mathrm{dist}(x, \mathcal{H}_w) = \frac{|\langle w, x\rangle|}{\|w\|}.$$

  If w separates the data, then $y_i\langle w, x_i\rangle > 0$ for all i, hence

$$\min_i \mathrm{dist}(x_i, \mathcal{H}_w) = \frac{\min_i y_i\langle w, x_i\rangle}{\|w\|}.$$

  Define
$$\mathcal{M}_S = \{w \in \mathbb{R}^d : y_i\langle w, x_i\rangle \geq 1,\ i = 1, \ldots, n\}.$$

  Let $w^* \in \mathcal{M}_S$ be the element of smallest norm. For any separating w, define

$$\gamma := \min_i y_i\langle w, x_i\rangle > 0, \quad \text{and} \quad \tilde{w} := \frac{w}{\gamma}.$$

  Then $\tilde{w} \in \mathcal{M}_S$, since
$$y_i\langle \tilde{w}, x_i\rangle = \frac{y_i\langle w, x_i\rangle}{\gamma} \geq 1.$$

  Hence
$$\min_i \mathrm{dist}(x_i, \mathcal{H}_w) = \frac{\gamma}{\|w\|} = \frac{1}{\|\tilde{w}\|}.$$

  Because w^* minimizes $\|w\|$ over $\mathcal{M}_S$,

$$\|w^*\| \leq \|\tilde{w}\|,$$

  which implies
$$\min_i \mathrm{dist}(x_i, \mathcal{H}_w) \leq \min_i \mathrm{dist}(x_i, \mathcal{H}_{w^*}).$$

  Thus, w^* achieves the *maximum margin* among all separating hyperplanes.

- Show that there are real numbers α_i such that $w^* = \sum_{i=1}^n \alpha_i x_i$.

  Consider the convex optimization problem:

$$\min_w \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad y_i\langle w, x_i\rangle \geq 1,\ i = 1, \dots, n.$$

  The Lagrangian is

$$\mathcal{L}(w, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^n \alpha_i(y_i\langle w, x_i\rangle - 1), \quad \alpha_i \geq 0.$$

  Setting the derivative with respect to w to zero (stationarity condition) yields:

$$\nabla_w \mathcal{L} = w - \sum_{i=1}^n \alpha_i y_i x_i = 0 \quad \Rightarrow \quad w^* = \sum_{i=1}^n \alpha_i y_i x_i.$$

  Thus, w^* lies in the span of the training examples $\{x_i\}$.

- Show that the α_i can be chosen so that $y_i\alpha_i$ are all nonnegative.

  Define $\tilde{\alpha}_i := \alpha_i y_i$. Then

$$w^* = \sum_{i=1}^n \tilde{\alpha}_i x_i, \quad \text{and} \quad y_i\tilde{\alpha}_i = y_i^2\alpha_i = \alpha_i \geq 0.$$

  Hence, the coefficients can be chosen so that $y_i\alpha_i \geq 0$ for all i.

2

Let u and v be D-dimensional unit vectors. Let M be a random matrix of dimension $d \times D$. Each entry of M is generated iid from a normal distribution with mean 0 and variance $1/d$.

1. Show that $\mathbb{E}[\langle Mu, Mv\rangle] = \langle u, v\rangle$.
2. Suppose $d \geq \frac{8}{\epsilon^2}$. Show that with probability at least $1 - e^{-1} - e^{-2}$,
$$\langle Mu, Mv\rangle \geq \langle u, v\rangle - \epsilon.$$
3. Now let's apply this to machine learning. Consider a set of n examples in D dimensional space that is linearly separable with margin y. That is, there are n examples, (x_i, y_i) with $y_i \in \{-1, 1\}$ and $||x_i|| \leq R$, and there is a unit vetor w so that $y_i\langle w, x_i\rangle \geq y$ for all i.

   Suppose that
$$d \geq 32\frac{R^2}{\gamma^2}\log(4n).$$
   Show that with probability at least $1/2$, $y_i\langle Mw, Mx_i\rangle \geq \frac{\gamma}{2}$ for all i. We can think of the vectors Mx_i as embeddings of the original data set in a lower dimensional space. This problem shows a random embedding already preserves much of the linear separability of data. An optimized embedding can do only better.
4. For parts 2 and 3, you can use the following fact about Gaussian random variables. If $g_1, \dots, g_k$ are independent Gaussian random variables with mean zero and variance 1, then
$$Pr\left[\frac{1}{m}\sum_{i=1}^m g_i^2 \geq 1 + \epsilon\right] \leq exp\left(-\frac{m\epsilon^2}{8}\right)$$
$$Pr\left[\frac{1}{m}\sum_{i=1}^m g_i^2 \geq 1 - \epsilon\right] \leq exp\left(-\frac{m\epsilon^2}{4}\right)$$

3

Consider the function $k : (0,1) \times (0,1) \to \mathbb{R}$ defined by $k(x_1, x_2) = min\{x_1, x_2\}$.

1. Prove that k is a valid kernel (Hint: write k as the integral of a product of two simple functions and then prove that its Gram matrices are positive semi-definite).

2. Now, consider a training set $\{(x_i, y_i)\}_{i=1,\ldots,n}$ with $y_i \in \mathbb{R}$ and distinct points x_i in $(0,1)$. Show that if we ran kernel regression without regularization on this data set, we would obtain zero training error. More precisely, find explicit coefficients α_j, in terms of the training data, such that for all points (x_i, y_i) in the training set we have

$$\sum_{j=1}^{n} \alpha_j \min\{x_j, x_i\} = y_i.$$