HW3

Kevin Chang

October 14, 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:

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

- Let w^* denote the element of $\mathcal{M}_S$ with smallest norm.

  Show that for any other w that separates the data

  $$\min dist_{1 \leq i \leq n}(x_i, \mathcal{H}_w) \leq \min_{1 \leq i \leq n} dist(x_i, \mathcal{H}_{w^*}).$$

  Proof. For any nonzero w, $\text{dist}(x, \mathcal{H}_w) = \frac{|\langle w, x \rangle|}{\|w\|}$. Let $\gamma_w := \min_i y_i \langle w, x_i \rangle > 0$ be the margin of w. Then

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

  Scale w to $\tilde{w} := \frac{w}{\gamma_w}$; then $\tilde{w} \in \mathcal{M}_S$ and $\|\tilde{w}\| = \frac{\|w\|}{\gamma_w}$. By optimality of w^* on $\mathcal{M}_S$, $\|w^*\| \leq \|\tilde{w}\| = \frac{\|w\|}{\gamma_w}$, hence

  $$\frac{\gamma_w}{\|w\|} \leq \frac{1}{\|w^*\|} = \min_i \text{dist}(x_i, \mathcal{H}_{w^*}),$$

  where the last equality uses $\min_i y_i \langle w^*, x_i \rangle = 1$ (if it were > 1, rescaling down would contradict minimality of $\|w^*\|$). This proves the claim. □

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

  Proof. Suppose, for the sake of contradiction, that w^* cannot be written as a linear combination of the training examples $\{x_i\}$. Then there exists a decomposition

  $$w^* = \sum_{i=1}^n \alpha_i x_i + v,$$

  where v is orthogonal to all x_i, i.e., $\langle v, x_i \rangle = 0$ for all i.

  Define $w' := \sum_{i=1}^n \alpha_i x_i$. Since $\langle v, x_i \rangle = 0$, we have

  $$y_i \langle w', x_i \rangle = y_i \langle w^*, x_i \rangle \geq 1,$$

  which implies $w' \in \mathcal{M}_S$.

  Moreover, $\|w'\| < \|w^*\|$ because $w^* = w' + v$ and $v \neq 0$ adds an orthogonal component, increasing the norm:

  $$\|w^*\|^2 = \|w'\|^2 + \|v\|^2 > \|w'\|^2.$$

  This contradicts the minimality of w^* as the smallest–norm element in $\mathcal{M}_S$. Hence, such a v cannot exist, and therefore

  $$w^* = \sum_{i=1}^n \alpha_i x_i.$$

  □

- 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 $\mathbb{E}[\langle Mu, Mv\rangle] = \langle u, v\rangle$.

   Answer:

$$\mathbb{E}\big[\langle Mu, Mv\rangle\big] = \mathbb{E}\big[u^\top M^\top Mv\big] = u^\top\, \mathbb{E}[M^\top M]\, v = u^\top I_D v = \langle u, v\rangle.$$

   (We used $\mathbb{E}[M^\top M] = I_D$ since each row of M has covariance $\frac{1}{d}I_D$ and there are d rows.)

2. Suppose $d \geq \frac{8}{\epsilon^2}$. Show that with probability at least $1 - e^{-1} - e^{-2}$,

$$\langle Mu, Mv\rangle \geq \langle u, v\rangle - \epsilon.$$

   Answer: Using the polarization identity,

$$\langle Mu, Mv\rangle = \tfrac{1}{4}\big(\|M(u+v)\|^2 - \|M(u-v)\|^2\big)\,.$$

   Let $z_\pm := u \pm v$. Then $\|Mz_\pm\|^2 = \frac{1}{d}\sum_{j=1}^{d} g_{j,\pm}^2$ with $g_{j,\pm} \sim \mathcal{N}(0, \|z_\pm\|^2)$ i.i.d., so $\frac{\|Mz_\pm\|^2}{\|z_\pm\|^2}$ is the average of d i.i.d. $\mathcal{N}(0,1)^2$ variables. By the bounds, for any $\varepsilon' > 0$,

$$\Pr\big[\big|\|Mz_+\|^2 - \|z_+\|^2\big| > \varepsilon'\|z_+\|^2\big] \leq e^{-d\varepsilon'^2/8},$$

$$\Pr\big[\big|\|Mz_-\|^2 - \|z_-\|^2\big| > \varepsilon'\|z_-\|^2\big] \leq e^{-d\varepsilon'^2/4}.$$

   On the intersection of these two events,

$$\big|\langle Mu, Mv\rangle - \langle u, v\rangle\big| = \tfrac{1}{4}\big|\big(\|Mz_+\|^2 - \|z_+\|^2\big) - \big(\|Mz_-\|^2 - \|z_-\|^2\big)\big| \leq \varepsilon'.$$

   Taking $\varepsilon' = \varepsilon$ and using the union bound gives

$$\Pr[\langle Mu, Mv\rangle \geq \langle u, v\rangle - \varepsilon] \geq 1 - e^{-d\varepsilon^2/8} - e^{-d\varepsilon^2/4}.$$

   If $d \geq 8/\varepsilon^2$, then $e^{-d\varepsilon^2/8} \leq e^{-1}$ and $e^{-d\varepsilon^2/4} \leq e^{-2}$, hence

$$\Pr[\langle Mu, Mv\rangle \geq \langle u, v\rangle - \varepsilon] \;\geq\; 1 - e^{-1} - e^{-2}.$$

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.

Answer: Assume a linearly separable dataset $\{(x_i, y_i)\}_{i=1}^n$ with $\|x_i\| \le R$, labels $y_i \in \{\pm 1\}$, and a unit vector w such that $y_i\langle w, x_i\rangle \ge \gamma$ for all i. Fix i and apply part (2) with the unit pair $u = w$ and $v = \frac{x_i}{\|x_i\|}$. With probability at least $1 - e^{-d\varepsilon^2/8} - e^{-d\varepsilon^2/4}$,

$$\langle Mw, M(x_i/\|x_i\|)\rangle \;\ge\; \langle w, x_i/\|x_i\|\rangle - \varepsilon.$$

Multiplying by $\|x_i\|$ and then by y_i yields

$$y_i\langle Mw, Mx_i\rangle \;\ge\; y_i\langle w, x_i\rangle - \|x_i\|\,\varepsilon \;\ge\; \gamma - R\varepsilon.$$

Choose $\varepsilon = \gamma/(2R)$. Then for this i,

$$\Pr[y_i\langle Mw, Mx_i\rangle \ge \gamma/2] \;\ge\; 1 - e^{-d\gamma^2/(32R^2)} - e^{-d\gamma^2/(16R^2)}.$$

If

$$d \;\ge\; 32\,\frac{R^2}{\gamma^2}\,\log(4n),$$

then $e^{-d\gamma^2/(32R^2)} \le \frac{1}{4n}$ and $e^{-d\gamma^2/(16R^2)} \le \frac{1}{4n}$, so for this i,

$$\Pr[y_i\langle Mw, Mx_i\rangle \ge \gamma/2] \;\ge\; 1 - \frac{1}{2n}.$$

Applying the union bound over all $i = 1, \ldots, n$,

$$\Pr[y_i\langle Mw, Mx_i\rangle \ge \gamma/2 \text{ for all } i] \;\ge\; 1 - n\cdot\frac{1}{2n} \;=\; \frac{1}{2}.$$

Thus, with probability at least $1/2$, every embedded example maintains margin at least $\gamma/2$ against Mw.

For parts 2 and 3, you can use the following fact about Gaussian random variables. If $g_1, \ldots, g_k$ are independent Gaussian random variables with mean zero and variance 1, then

$$Pr\left[\frac{1}{m}\sum_{i=1}^{m} g_i^2 \ge 1 + \epsilon\right] \le exp\left(-\frac{m\epsilon^2}{8}\right)$$

$$Pr\left[\frac{1}{m}\sum_{i=1}^{m} g_i^2 \ge 1 - \epsilon\right] \le exp\left(-\frac{m\epsilon^2}{4}\right)$$

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).

   Answer: For $x \in (0,1)$ define the feature map $\phi_x : [0,1] \to \mathbb{R}$ by

   $$\phi_x(t) := \mathbf{1}\{t \le x\}.$$

   Then

   $$k(x_1, x_2) = \int_0^1 \phi_{x_1}(t)\,\phi_{x_2}(t)\,dt = \langle \phi_{x_1}, \phi_{x_2}\rangle_{L^2[0,1]}.$$

   Hence k is an inner-product kernel.

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

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

Answer: Without loss of generality, assume the inputs are sorted:

$$0 < x_1 < x_2 < \cdots < x_n < 1.$$

Define

$$f(x) = \sum_{j=1}^{n} \alpha_j \min\{x, x_j\}.$$

Then f is continuous and piecewise linear, with slope on each interval $(x_{i-1}, x_i]$ given by

$$f'(x) = \sum_{j:\, x_j \geq x} \alpha_j.$$

We want $f(x_i) = y_i$ for all i. To achieve this, define the slopes between adjacent points:

$$\beta_1 = \frac{y_1}{x_1}, \qquad \beta_i = \frac{y_i - y_{i-1}}{x_i - x_{i-1}} \quad \text{for } i = 2, \ldots, n.$$

These β_i describe the desired piecewise-linear interpolant through $(0, 0)$ and (x_i, y_i).
Since $f'(x) = \sum_{j=i}^{n} \alpha_j = \beta_i$ for $x \in (x_{i-1}, x_i]$, we can recover α from the backward differences:

$$\alpha_i = \beta_i - \beta_{i+1} \quad (i = 1, \ldots, n-1), \qquad \alpha_n = \beta_n,$$

where we take $\beta_{n+1} := 0$.
Then, for each i,

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

so the regression interpolates the data exactly—yielding zero training error.

4

In the high-dimensional problems, there are usually an infinite number of possible models that perfectly fit the observed data. When a problem has multiple solutions, different optimization algorithms can find entirely different solutions to the same problem. Even though all of the solutions perfectly fit the training data, their generalization performance can be vastly different. In this problem, we explore this phenomenon for two widely used optimization algorithms: gradient descent and Adam. Consider a linear, binary classification problem under the squared loss. Let $X \in \mathbb{R}^{n \times d}$ be an $n \times d$ matrix of features, $y \in \{-1, 1\}^n$ be the corresponding vector of labels, and $\theta \in \mathbb{R}^d$ be the parameter vector. We wish to minimize the empirical risk

$$R_S[\theta] = \frac{1}{2}||X\theta - y||_2^2$$

Assume that the rows of X are linearly independent and that $d > n$.