CIS5200: Machine Learning Spring 2025

Homework 1

Release Date: January 30, 2025 *Due Date: February 13, 2025*

Problem 1: Margin Perceptron

1.1 Let's prove the growth lemma.

Proof. On a mistake round or when $|w_t^\top x_i| \leq 1$:

$$\begin{aligned} w_*^\top w_{t+1} &= w_*^\top (w_t + y_i x_i) \\ &= w_*^\top w_t + y_i w_*^\top x_i \\ &\geq w_*^\top w_t + \gamma \end{aligned}$$

The last inequality uses our assumption that $y_i w_*^\top x_i \geq \gamma$ for all i, which follows from the fact that w_* correctly classifies all points with margin at least γ. □

1.2 Let's prove the control lemma.

Proof. On an update, we have either:

Case 1 (mistake): $y_i \neq \text{sign}(w_t^\top x_i)$, which means $y_i w_t^\top x_i \leq 0$. This occurs when the current prediction is wrong.

Case 2 (margin violation): $|w_t^\top x_i| \leq 1$. This occurs when the prediction is correct but not confident enough.

This implies that $y_i w_t^\top x_i \leq 2$ in both cases. This gives us:

$$\begin{aligned} \|w_{t+1}\|_2^2 &= \|w_t + y_i x_i\|_2^2 \\ &= \|w_t\|_2^2 + 2 y_i w_t^\top x_i + \|x_i\|_2^2 \\ &\leq \|w_t\|_2^2 + 2 \cdot 1 + 1 \\ &= \|w_t\|_2^2 + 3 \end{aligned}$$

□

1.3 Let's combine the lemmas.

Proof. From the growth lemma after T rounds:

$$w_*^\top w_{T+1} \geq \gamma T$$

By Cauchy-Schwarz (since $\|w_*\|_2 = 1$ by assumption):

$$\gamma T \leq w_*^\top w_{T+1} \leq \|w_*\|_2 \|w_{T+1}\|_2 = \|w_{T+1}\|_2$$

From the control lemma after T rounds:

$$\|w_{T+1}\|_2^2 \leq 3T$$

Therefore:

$$\gamma T \leq \|w_{T+1}\|_2 \leq \sqrt{3T}$$

This elegant combination shows that while the weight vector's length grows as $\sqrt{T}$, its alignment with w_* grows linearly with T, forcing convergence. □

1.4 Convergence bound.

Proof. From 1.3:

$$\gamma T \leq \sqrt{3T} \implies \gamma^2 T^2 \leq 3T \implies T \leq \frac{3}{\gamma^2}$$

□

1.5 Let's prove the margin bound.

Proof. At termination, for all i:

$$|w^\top x_i| > 1 \text{ and } y_i w^\top x_i > 0$$

Therefore, the normalized margin (distance from point to hyperplane) is:

$$\frac{|w^\top x_i|}{\|w\|_2} > \frac{1}{\|w\|_2} \geq \frac{1}{\sqrt{3T}} \geq \frac{\gamma}{3}$$

where we used: (1) the control lemma bound $\|w\|_2 \leq \sqrt{3T}$ (2) the convergence bound $T \leq \frac{3}{\gamma^2}$. This shows that the final classifier achieves a margin proportional to the optimal margin γ. □

1.6 Benefits of margin: A large margin provides robustness to noise and better generalization, as small perturbations to inputs are less likely to change predictions when the decision boundary is far from training points.

Problem 2: Bayes Optimal Classifier and Squared Loss

2.1 Let's find $h^*(x)$.

Proof. The expected squared loss for fixed x is:

$$\mathop{\mathbb{E}}_{y|x}[(h(x)-y)^2] = \eta(x)(h(x)-1)^2 + (1-\eta(x))(h(x)+1)^2$$

To get $h^*(x)$, we take the derivative with respect to $h(x)$ and set it to zero:

$$\begin{aligned}
2\eta(x)(h^*(x)-1) + 2(1-\eta(x))(h^*(x)+1) &= 0 \\
2\eta(x)h^*(x) - 2\eta(x) + 2h^*(x) - 2\eta(x)h^*(x) + 2 - 2\eta(x) &= 0 \\
2h^*(x) - 4\eta(x) + 2 &= 0 \\
h^*(x) &= 2\eta(x) - 1
\end{aligned}$$

This minimizes the expected squared loss since the second derivative is positive. □

2.2 Let's derive $\eta(x)$.

Proof. By Bayes rule:

$$\begin{aligned}
\eta(x) = P(y=1|x) &= \frac{P(x|y=1)P(y=1)}{P(x|y=1)P(y=1) + P(x|y=-1)P(y=-1)} \\
&= \frac{\frac{1}{2}\exp(-\frac{1}{2}\|x-\mu\|^2)}{(\frac{1}{2}\exp(-\frac{1}{2}\|x-\mu\|^2) + \frac{1}{2}\exp(-\frac{1}{2}\|x+\mu\|^2))}
\end{aligned}$$

Writing out the squared norms:

$$\begin{aligned}
\|x-\mu\|^2 &= \|x\|^2 - 2\mu^\top x + \|\mu\|^2 \\
\|x+\mu\|^2 &= \|x\|^2 + 2\mu^\top x + \|\mu\|^2
\end{aligned}$$

Therefore:

$$\begin{aligned}
\eta(x) &= \frac{\exp(-\frac{1}{2}(\|x\|^2 - 2\mu^\top x + \|\mu\|^2))}{\exp(-\frac{1}{2}(\|x\|^2 - 2\mu^\top x + \|\mu\|^2)) + \exp(-\frac{1}{2}(\|x\|^2 + 2\mu^\top x + \|\mu\|^2))} \\
&= \frac{\exp(\mu^\top x)}{\exp(\mu^\top x) + \exp(-\mu^\top x)} \\
&= \frac{1}{1+\exp(-2\mu^\top x)}
\end{aligned}$$

□

2.3 Let's find the decision boundary.

Proof. From 2.1 and 2.2:

$$h^*(x) = 2\eta(x) - 1 = \frac{2}{1+\exp(-2\mu^\top x)} - 1 = \tanh(\mu^\top x)$$

Since $\tanh(z) > 0$ if and only if $z > 0$:

$$h^*(x) > 0 \iff \mu^\top x > 0$$

Therefore $w = \mu$ and $b = 0$ give the linear decision boundary $w^\top x + b = 0$. □

Problem 3: k-NN Analysis

3.1 Let's analyze distance changes.

Proof. Let x and x' differ only in coordinate j by ϵ (i.e., $x'_j = x_j + \epsilon$ and $x'_i = x_i$ for all $i \neq j$). For any training point z:

$$\begin{aligned}
|\|x - z\|_2 - \|x' - z\|_2| &\leq \|x - z - (x' - z)\|_2 \text{ (reverse triangle inequality)} \\
&= \|x - x'\|_2 \\
&= \sqrt{\sum_{i=1}^{d} (x_i - x'_i)^2} \\
&= \sqrt{(x_j - x'_j)^2} \text{ (since only coordinate } j \text{ differs)} \\
&= |x_j - x'_j| = |\epsilon|
\end{aligned}$$

This shows that if we perturb one coordinate by ϵ, the distance to any training point changes by at most $|\epsilon|$, demonstrating local stability of distances. $\square$

3.2 Let's prove stability.

Proof. Let x' be the perturbed point where each coordinate differs from x by at most ϵ. For any training point z, we can bound the change in distance in two ways:

By applying the result from 3.1 sequentially to each coordinate change:

$$|\|x - z\|_2 - \|x' - z\|_2| \leq d\epsilon \text{ (since each coordinate contributes at most } \epsilon)$$

Setting $\epsilon = \frac{\Delta}{2d}$ ensures total change $\leq \frac{\Delta}{2}$. Let z_1, z_2 be the nearest and second-nearest neighbors to x. Then:

$$\begin{aligned}
\|x' - z_1\|_2 &\leq \|x - z_1\|_2 + \frac{\Delta}{2} \\
\|x' - z_2\|_2 &\geq \|x - z_2\|_2 - \frac{\Delta}{2} = \|x - z_1\|_2 + \frac{\Delta}{2}
\end{aligned}$$

Therefore z_1 remains the nearest neighbor to x', preserving the 1-NN prediction. $\square$

Actually you can get a stronger bound using the direct L2 norm calculation:

$$\begin{aligned}
|\|x - z\|_2 - \|x' - z\|_2| &\leq \|x - x'\|_2 \text{ (reverse triangle inequality)} \\
&= \sqrt{\sum_{i=1}^{d} (x_i - x'_i)^2} \\
&\leq \sqrt{\sum_{i=1}^{d} \epsilon^2} = \epsilon\sqrt{d}.
\end{aligned}$$

So only $\epsilon = \frac{\Delta}{2\sqrt{d}}$ suffices to ensure stability.