CSCI567 HW1

Angzhan He, Gaoyuan Jiang

February 3, 2024

1 1.1

$$\begin{aligned} w_{k+1}^T w_{opt} &= (w_k + y_i x_i)^T w_{opt} \\ &= w_k^T w_{opt} + y_i x_i^T w_{opt} \\ &= w_k^T w_{opt} + y_i w_{opt}^T x_i \\ &\geq w_k^T w_{opt} + \gamma ||w_{opt}||_2 \end{aligned}$$

1.2

$$\begin{aligned} w_{k+1}^T w_{k+1} &= (w_k + y_i x_i)^T (w_k + y_i x_i) \\ &= w_k^T w_k + w_k^T y_i x_i + y_i x_i^T w_k + y_i^2 x_i^T x_i \\ &= ||w_k||_2^2 + 2 y_i w_k^T x_i + y_i^2 ||x_i||_2^2 \\ &\leq ||w_k||_2^2 + y_i^2 R^2 \\ &\leq ||w_k||_2^2 + R^2 \end{aligned}$$

1.3

$$\begin{gathered} w_{k+1}^T w_{opt} \geq w_k^T w_{opt} + \gamma ||w_{opt}||_2 \\ M updates \\ w_{k+1}^T w_{opt} \geq \gamma M \\ ||w_{k+1}||_2 ||w_{opt}||_2 \geq w_{k+1}^T w_{opt} \geq \gamma M \\ ||w_{k+1}||_2 \geq \gamma M \end{gathered}$$

$$\begin{gathered} ||w_{k+1}||_2^2 \leq ||w_k||_2^2 + R^2 \\ M updates \\ ||w_{k+1}||_2^2 \leq R^2 M \\ ||w_{k+1}||_2 \leq R\sqrt{M} \end{gathered}$$

1.4

$$\begin{gathered} \gamma M \leq ||w_{k+1}||_2 \leq R\sqrt{M} \\ \gamma^2 M^2 \leq R^2 M \\ M \leq R^2/\gamma^2 \end{gathered}$$

2.1

Algorithm: For all positive data points in the training set, find the smallest and the largest values of x_1 and x_2, which correspond to a_1, b_1, a_2 and b_2 respectively.

Proof: The realizable assumption shows that there exists a rectangle B^* that perfectly classifies the training data. The rectangle B_S we get by the algorithm have an empirical risk of 0, which is the minimum possible. Thus, the rectangle B_S is an empirical risk minimizer.

2.2

From a probabilistic perspective and with respect to $0-1$ loss, $R(f_{S'}^{ERM}) \geq 0.5$ indicates the misclassifying probability is greater than 0.5. If we need to let this model classify every data points in training set $\{(\mathbf{x}, y), i \in [n]\}$ correctly, it means we can not select any data point from $B^* - B_{S'}$. The probability mass (with respect to D) of $B^* - B_{S'}$ is larger than 0.5. If we draw data point i.i.d from distribution D, then the possibility of selecting such a bad training set of size n is less than 0.5^n, which is non-zero, but very small when n is large enough.

2.3

Step1: According to the definition of empirical risk minimizer and realizability assumption, B_S must be contained within B^* to have zero empirical risk.
Step2: B_i has a probability mass of $\varepsilon/4$ by construction, since there are four such rectangles, the combined B_S where f_S^{ERM} could potentially fail to classify correctly is less than $4 \times \varepsilon/4 = \varepsilon$.
Step3: The probability that none of the n examples in S are in B_i is $(1-\varepsilon/4)^n$.

$$P = (1-\varepsilon/4)^n$$
$$log(P) = nlog(1-\varepsilon/4)$$
$$log(P) \leq -n(\varepsilon/4)$$
$$log(P) \leq log(\delta/4)$$
$$P \leq \delta/4$$

Step4: The union bound states that the probability of at least one of a set of events occurring is on greater than the sum of the probabilities of the individual events. Apply this to the probability that S does not contain an example from each B_i, the sum of the probability is $4 \times (\delta/4) = \delta$. Thus, the probability that S contains all examples from each B_i is at least $1-\delta$.

2.4

In R^d, define $2d$ critical regions (similar to the B_i rectangles from the 2-dimensional case) surrounding the true B^*, each with a probability mass of $\varepsilon/(2d)$ with respect to the distribution D.
If S contains positive examples in all of the critical regions, then $R(f_S^{ERM}) \leq (2d) \times (\varepsilon/(2d)) = \varepsilon$.
According to the union bound, for each of the $2d$ critical regions, we want the probability that the sample S does not contain a positive example from that region to be less than $\delta/(2d)$.

$$P = (1-\varepsilon/(2d))^n$$
$$log(P) = nlog(1-\varepsilon/(2d))$$
$$log(P) \leq -n(\varepsilon/(2d))$$

To make $P \leq \delta/(2d)$, we get $n \geq \frac{2dlog(2d/\delta)}{\varepsilon}$.