CSE 546: Machine Learning **Lecture 9**

Online Learning & Margins

Instructor: Sham Kakade

1 Introduction

There are two common models of study:

Online Learning No assumptions about data generating process. Worst case analysis. Fundamental connections to Game Theory.

Statistical Learning Assume data consists of independently and identically distributed examples drawn according to some fixed but *unknown* distribution.

Our examples will come from some space $\mathcal{X} \times \mathcal{Y}$. Given a *data set*

$$\{(x_t, y_t)\}_{t=1}^T \in (\mathcal{X} \times \mathcal{Y})^T ,$$

our goal is to predict y_{T+1} for a new point x_{T+1}. A *hypothesis* is simply a function $h : \mathcal{X} \to \mathcal{Y}$. Sometimes, a hypothesis will map to a set $\mathcal{D}$ (for decision space) larger than $\mathcal{Y}$. Depending on the nature of the set $\mathcal{Y}$, we get special cases of the general prediction problem. Here, we examine the case of binary classification where $\mathcal{Y} = \{-1, +1\}$.

A set of hypotheses is often called a *hypotheses class*.

In the online learning model, learning proceeds in rounds, as we see examples one by one. Suppose $\mathcal{Y} = \{-1, +1\}$. At the beginning of round t, the learning algorithm $\mathcal{A}$ has the hypothesis h_t. In round t, we see x_t and predict $h_t(x_t)$. At the end of the round, y_t is revealed and $\mathcal{A}$ makes a mistake if $h_t(x_t) \neq y_t$. The algorithm then updates its hypothesis to h_{t+1} and this continues till time T.

Suppose the labels were actually produced by some function f in a given hypothesis class $\mathcal{C}$. Then it is natural to bound the total number of mistakes the learner commits, no matter how long the sequence. To this end, define

$$\text{mistake}(\mathcal{A}, \mathcal{C}) := \max_{f \in \mathcal{C}, T, x_{1:T}} \sum_{t=1}^T \mathbf{1}\left[h_t(x_t) \neq f(x_t)\right] .$$

2 Linear Classifiers and Margins

Let us now look at a concrete example of a hypothesis class. Suppose $\mathcal{X} = \mathbb{R}^d$ and we have a vector $w \in \mathbb{R}^d$. We define the hypothesis,

$$h_w(x) = \text{sgn}(w \cdot x) ,$$

where $\text{sgn}(z) = 1$ if z is positive and -1 otherwise. With some abuse of terminology, we will often speak of “the hypothesis w” when we actually mean “the hypothesis h_w”. The class of *linear classifiers* in the (uncountable) hypothesis class

$$\mathcal{C}_{\text{lin}} := \left\{ h_w \,\middle|\, w \in \mathbb{R}^d \right\} .$$

Note that w and αw yield the same linear classifier for any scalar $\alpha > 0$.

Suppose we have a data set that is *linearly separable*. That is, there is a w_* such that,

$$\forall t \in [T],\ y_t = \mathrm{sgn}(w_* \cdot x_t)\ . \tag{1}$$

Separability means that $y_t(w_* \cdot x_t) > 0$ for all t. The minimum value of this quantity over the data set is referred to as the *margin*. Let us make the assumption that the margin is lower bounded by 1.

Assumption M. *(Margin of* 1*) Without loss of generality suppose* $\|x_t\| \leq 1$*. Suppose there exists a* $w_* \in \mathbb{R}^d$ *for which* (1) *holds. Further assume that*

$$\min_{t\in[T]} y_t(w_* \cdot x_t) \geq 1\ , \tag{2}$$

Note the choice of 1 *is arbitrary.*

Note that the above implies that:

$$\min_{t\in[T]} y_t(\frac{w_*}{\|w_*\|} \cdot x_t) \geq \frac{1}{\|w_*\|}\ .$$

In other words, the width of the strip separating the positives from the negatives is of size $\frac{2}{\|w_*\|}$. Sometimes the margin is define this way (where we assume that instead $\|w_*\| = 1$ and that the margin is some positive value rather than 1).

2.1 The Perceptron Algorithm

Algorithm 1 PERCEPTRON

$w_1 \leftarrow \mathbf{0}$
for $t = 1$ to T **do**
  Receive $x_t \in \mathbb{R}^d$
  Predict $\mathrm{sgn}(w_t \cdot x_t)$
  Receive $y_t \in \{-1, +1\}$
  if $\mathrm{sgn}(w_t \cdot x_t) \neq y_t$ **then**
    $w_{t+1} \leftarrow w_t + y_t x_t$
  else
    $w_{t+1} \leftarrow w_t$
  end if
end for

The following theorem gives a dimension independent bound on the number of mistakes the PERCEPTRON algorithm makes.

Theorem 2.1. *Suppose Assumption M holds. Let*

$$M_T := \sum_{t=1}^{T} \mathbf{1}\left[\mathrm{sgn}(w_t \cdot x_t) \neq y_t\right]$$

denote the number of mistakes the PERCEPTRON *algorithm makes. Then we have,*

$$M_T \leq \|w_*\|^2\ .$$

Second, if we had instead assumed that $\|x_t\| \leq X_+$, then the above would be:

$$M_T \leq \cdot X_+^2 \|w_*\|^2\ .$$

Proof. Define $m_t = 1$ if a mistake occurs at time t and 0 otherwise. We have that:

$$w_{t+1} = w_t + m_t y_t x_t$$

Now observe that:

$$\begin{aligned}
\|w_{t+1} - w_*\|^2 &= \|w_t + m_t y_t x_t - w_*\|^2 \\
&= \|w_t - w_*\|^2 + 2m_t y_t x_t (w_t - w_*) + m_t^2 y_t^2 \|x_t\|^2 \\
&= \|w_t - w_*\|^2 + 2m_t y_t x_t (w_t - w_*) + m_t \|x_t\|^2 \\
&\le \|w_t - w_*\|^2 + 2m_t y_t x_t (w_t - w_*) + m_t \\
&\le \|w_t - w_*\|^2 - 2m_t + m_t \\
&\le \|w_t - w_*\|^2 - m_t
\end{aligned}$$

where the second to last step holds since we have that:

$$m_t y_t x_t (w_t - w_*) \le m_t y_t x_t w_t - m_t < -m_t$$

using the margin assumption and that $y_t x_t w_t < 0$ when there is a mistake.

Hence, we have that:

$$m_t \le \|w_t - w_*\|^2 - \|w_{t+1} - w_*\|^2$$

This implies:

$$M_T = \sum_{t=1}^{T} m_t \le \|w_1 - w_*\|^2 - \|w_{T+1} - w_*\|^2 \le \|w_*\|^2$$

which completes the proof. □

3 SVMs

The SVM loss function can be viewed as a relaxation to the classification loss. The *hinge* loss on a pair (x, y) is defined as:

$$\ell((x, y), w) = \max\{0, 1 - yw^\top x\}$$

In other words, we penalize with a linear loss when $yw^\top x$ is 1 or less. Note that we could actually penalize when we have a correct prediction (if $0 \le yw^\top x \le 1$ then our prediction is correct and we are still penalized). In this latter case, we call this a 'margin' mistake.

Note that the gradient of this loss is:

$$\nabla \ell((x, y), w) = -yx \text{ if } yw^\top x < 1$$

and the gradient is 0 otherwise.

The SVM seeks to minimize the following objective:

$$\frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i w^\top x_i\} + \lambda \|w\|^2$$

As usual, the algorithm can be kernelized.