

Lecture 2 — January 26

Lecturer: Martin Wainwright

Scribe: Aditi Shrikumar

Note: These lecture notes are still rough, and have only have been mildly proofread.

2.1 Announcements

- HW1 posted, due 2/9
- Scribe materials on web page

Today:

- Binary classification
- Perceptron and max-margin

2.2 Last Time: binary classification, loss, and risk

We start with features $X \in \mathcal{X}$ and labels $Y \in \mathcal{Y}$ where (X, Y) are distributed according to some unknown \mathbb{P} . The goal is to find some decision rule $f : \mathcal{X} \rightarrow \mathcal{Y}$

Given a decision rule $f : \mathcal{X} \rightarrow \mathcal{Y}$, its 0-1 loss $l(y, f(x))$ is:

$$l(y, f(x)) = \mathbb{I}[y \neq f(x)] \quad (2.1)$$

in other words,

$$l(y, f(x)) = \begin{cases} 1 & \text{if } y \neq f(x) \\ 0 & \text{otherwise} \end{cases} . \quad (2.2)$$

And its risk $\mathbb{R}(f)$ is:

$$\mathbb{R}(f) = \mathbb{E}[l(Y, f(X))] = \mathbb{P}(Y \neq f(X)) \quad (2.3)$$

Under 0-1 loss, the risk is equal to the probability of misclassification.

2.3 Today

2.3.1 The optimal decision function

What is the optimal decision function? What is the rule $g^* : \mathcal{X} \rightarrow \mathcal{Y}$ such that $\mathbb{R}(g^*) = \mathbb{P}(y \neq g^*(x)) < \mathbb{P}(y \neq g(x)) = \mathbb{R}(g)$ for all other rules $g : \mathcal{X} \rightarrow \mathcal{Y}$?

\mathbb{P} is unknown in “real life”, but let’s assume that it is known for now. Let us parametrize \mathbb{P} over (X, Y) by $\mu(A) = \mathbb{P}(X \in A)$ (the marginal over features), and $\eta(x) = \mathbb{P}(Y = +1|X = x)$.

Define the *Bayes Decision Rule* $g^*(x)$:

$$g^*(x) = \begin{cases} +1 & \text{if } \eta(x) \geq 1/2 \\ -1 & \text{otherwise} \end{cases} . \quad (2.4)$$

Claim: This is the optimal rule, i.e. all other rules g have $\mathbb{R}(g) \geq \mathbb{R}(g^*)$ **Note:** This is equivalent to the likelihood ratio test:

$$\eta(x) \geq 1/2 \Leftrightarrow \frac{\mathbb{P}(Y = +1|X = x)}{\mathbb{P}(Y = -1|X = x)} \geq 1. \quad (2.5)$$

Theorem 2.1. For any decision function $g : \mathcal{X} \rightarrow \{+1, -1\}$ we have $\mathbb{R}(g^*) \leq \mathbb{R}(g)$.

Proof: For any fixed $x \in \mathcal{X}$, and any g :

$$\mathbb{P}(Y \neq g(X)|X = x) = 1 - \{\mathbb{P}(Y = 1, g(X) = 1|X = 1) + \mathbb{P}(Y = -1, g(X) = -1|X = x)\} \quad (2.6)$$

$$= 1 - \{\mathbb{E}[\mathbb{I}(g(x = 1))]\eta(x) + \mathbb{E}[\mathbb{I}(g(x = -1))\eta(x)]\}. \quad (2.7)$$

Hence,

$$\mathbb{P}(Y \neq g(X)|X = x) - \mathbb{P}(Y \neq g^*(X)|X = x) = \eta(x)\{\mathbb{E}[\mathbb{I}(g^*(x) = 1)] - \mathbb{E}[\mathbb{I}(g(x) = 1)]\} + (1 - \eta(x))\{\mathbb{E}[\mathbb{I}(g^*(x) = -1)] + \mathbb{E}[\mathbb{I}(g(x) = -1)]\} \quad (2.8)$$

$$= (2\eta(x) - 1)\{\mathbb{E}[\mathbb{I}(g^*(x) = 1)] - \mathbb{E}[\mathbb{I}(g(x) = 1)]\} \geq 0 \quad (2.9)$$

(case-by-case). Taking expectation w.r.t X implies $\mathbb{R}(g) - \mathbb{R}(g^*) \geq 0$ by definition of g^* . \square

2.3.2 Example: predicting pass/fail in a class

Say the features we have are:

$$X_1 = \text{number of hours spent sleeping per day} \quad (2.10)$$

$$X_2 = \text{number of beers per day} \quad (2.11)$$

$$X_3 = \text{“laziness” – unobservable} \quad (2.12)$$

And we want to predict whether the student with those features passes a class.

$$Y = \begin{cases} +1 & \text{(pass) if } x_1 + x_2 + x_3 \leq 7 \\ -1 & \text{otherwise} \end{cases}. \quad (2.13)$$

This defines $\eta(x)$ implicitly. We want to find a decision rule based only on (X_1, X_2) .

Say the X_i 's are distributed according to $\text{Exp}[1]$, i.i.d. We can compute the conditional probability

$$\eta(x_1, x_2) = \mathbb{P}(Y = +1|X_1 = x_1, X_2 = x_2) \quad (2.14)$$

$$= \text{(some steps... left as an exercise)} \quad (2.15)$$

$$= \max(0, 1 - e^{-(7-x_1-x_2)}) \quad (2.16)$$

This gives us the bayes decision rule g^* :

$$g^* = \begin{cases} +1 & \text{if } x_1 + x_2 \leq 7 - \log(2) \\ -1 & \text{otherwise} \end{cases}. \quad (2.17)$$

2.3.3 Plug-in rules

In practice, we don't know $\eta(\cdot)$, so we cannot compute the Bayes rule g^* . But perhaps you can use data to estimate η and apply plug-in principle:

$$g_{\hat{\eta}}(x) = \begin{cases} +1 & \text{if } \hat{\eta}(x) \geq 1/2 \\ -1 & \text{otherwise} \end{cases}. \quad (2.18)$$

Theorem 2.2. For any plug in rule, the excess risk $\mathbb{R}(g_{\hat{\eta}}) - \mathbb{R}(g^*) < 2|\mathbb{E}[\eta^*(x) - \hat{\eta}(x)]|$

Proof: If $g^*(x) = g_{\hat{\eta}}(x)$, then $\mathbb{P}(g_{\hat{\eta}}(x) \neq Y|X = x) - \mathbb{P}(g^*(x) \neq Y|X = x) = 0$. Otherwise, when $g_{\hat{\eta}}(x) \neq g^*(x)$, from theorem 2.1:

$$\mathbb{P}(g_{\hat{\eta}}(x) \neq Y|X = x) - \mathbb{P}(g^*(x) \neq Y|X = x) \quad (2.19)$$

$$= (2\eta(x) - 1)\mathbb{E}[\mathbb{I}(g^*(x) = 1) - \mathbb{I}(g_{\hat{\eta}}(x) = 1)] \quad (2.20)$$

$$= |2\eta(x) - 1|\mathbb{E}[\mathbb{I}(g^*(x) \neq g_{\hat{\eta}}(x))] \quad (2.21)$$

Taking expectations, this means that

$$\mathbb{R}(g_{\hat{\eta}}) - \mathbb{R}(g^*) < 2 \mathbb{E}|\eta^*(x) - \hat{\eta}(x)| \quad (2.22)$$

If $g^*(x) \neq g_{\hat{\eta}}(x)$ then $|\eta(x) - \hat{\eta}(x)| \geq |\eta(x) - 1/2|$. Result follows. \square

2.3.4 The Perceptron Algorithm

The perceptron algorithm searches over linear threshold functions

$$\mathcal{F} = \{\text{sgn}[f_{\theta}(\cdot)] \mid f_{\theta}(x) = \langle \theta, x \rangle, \theta \in \mathbb{R}^d\} \quad (2.23)$$

Given samples $\mathcal{D}_n = \{(x^{(i)}, y^{(i)}) \in \mathcal{X} \times \{+1, -1\}, i = 1, \dots, n\}$, and given weight vector $\theta \in \mathbb{R}^d$, define the set of mistakes $m(\theta) = \{i \in \{1, \dots, n\} : y^{(i)} \langle \theta, x^{(i)} \rangle < 0\}$..

The perceptron algorithm generates a sequence of vectors $\theta^0, \theta^1, \theta^2, \dots$

1. Initialize $\theta^0 = 0$.
2. For $t = 0, 1, 2, \dots$, while $m(\theta^t) \neq \emptyset$, pick some $i \in m(\theta^t)$ and update $\theta^{t+1} = \theta^t + y^{(i)}x^{(i)}$.

This algorithm needs classes that are separated by lines.

Definition: A data set is linearly separable if there exists a $\theta \in \mathbb{R}^d$ such that $m(\theta) = \emptyset$ i.e. there is a line between the two classes that makes no mistakes.

Theorem 2.3. For any linearly separable data set \mathcal{D}_n , the perceptron algorithm terminates in at most $T = R^2/\delta^2$ steps where

$$R = \max_{i=1, \dots, n} \|x^{(i)}\|_2 \quad (2.24)$$

$$\delta = \min_{i=1, \dots, n} \left(\frac{y^{(i)} \langle \theta^*, x^{(i)} \rangle}{\|\theta^*\|_2} \right) > 0 \text{ for some } \theta^* \text{ with } m(\theta^*) = \emptyset \quad (2.25)$$

δ is called the margin. It is the distance between the line θ , and the closest data point to that line.

Proof: Say we make a mistake on sample i at step t . Then

$$\langle \theta^*, \theta^{t+1} \rangle = \langle \theta^*, \theta^t \rangle + y^{(i)} \langle \theta^*, x^{(i)} \rangle \quad (2.26)$$

$$\geq \langle \theta^*, \theta^t \rangle + \delta \|\theta^*\|_2 \text{ (by definition of } \delta\text{)}. \quad (2.27)$$

Hence, by induction, we get that

$$\langle \theta^*, \theta^t \rangle \geq t\delta \|\theta^*\|_2 \quad (2.28)$$

On the other hand,

$$\|\theta^{t+1}\|_2^2 = \|\theta^t\|_2^2 + \|x^{(i)}\|_2^2 + 2y^{(i)} \langle \theta^t, x^{(i)} \rangle \quad (2.29)$$

$$\leq \|\theta^t\|_2^2 + R^2, \text{ because } y^{(i)} \langle \theta^t, x^{(i)} \rangle < 0 \text{ (mistake)} \quad (2.30)$$

$$\implies \|\theta^t\|_2^2 \leq tR^2, \text{ by induction} \quad (2.31)$$

Hence:

$$\delta t \|\theta^*\|_2 \leq \langle \theta^*, \theta^t \rangle \quad (2.32)$$

$$\leq \|\theta^*\|_2 \|\theta^t\|_2 \quad (2.33)$$

$$\leq \|\theta^*\|_2 \sqrt{t} R \quad (2.34)$$

$$\implies t \leq \frac{R^2}{\delta^2} \quad (2.35)$$

\square