


4. Linear threshold functions.

5. Perceptron algorithm.
Assume:

- There is a probability distribution $P$ on $\mathcal{X} \times \mathcal{Y}$,
- The pairs $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y)$ are chosen independently according to $P$

The aim is to choose $f$ with small risk:

$$R(f) = \mathbb{E} \ell(f(X), Y).$$

If we choose $f \in F$, can we achieve small excess risk,

$$R(f_n) - \inf_{f \in F} R(f)?$$
Consider two-class classification: $\mathcal{Y} = \{\pm 1\}$.

Notation: represent the joint distribution $P$ on $\mathcal{X} \times \mathcal{Y}$ as the pair $(\mu, \eta)$, where $\mu$ is the marginal distribution on $\mathcal{X}$ and $\eta$ is the conditional probability of $Y$ given $X$,

$$\eta(x) = P(Y = 1|X = x).$$
Pattern classification

If we know $\eta$, we could use it to find a decision rule that minimizes risk. To see this, notice that we can write the expected loss as an expectation of a conditional expectation,

$$R(f) = E\ell(f(X), Y)$$

$$= EE[\ell(f(X), Y)|X]$$

$$= E(\ell(f(X), 1)P(Y = 1|X) + \ell(f(X), -1)P(Y = -1|X))$$

$$= E(1[f(X) \neq 1]\eta(X) + 1[f(X) \neq -1](1 - \eta(X)))$$

$$= E(1[f(X) \neq 1]\eta(X) + (1 - 1[f(X) \neq 1])(1 - \eta(X)))$$

$$= E(1[f(X) \neq 1](2\eta(X) - 1) + 1 - \eta(X)).$$
Clearly, this expectation is minimized by choosing \( f = f^* \), where

\[
f^*(x) = \begin{cases} 
1 & \text{if } \eta(x) \geq 1/2, \\
-1 & \text{if } \eta(x) < 1/2.
\end{cases}
\]

Obviously, if \( \eta(x) = 1/2 \), the choice does not affect the risk.

Denote the optimal risk (the *Bayes risk*), by

\[
R^* = \inf_f R(f) = R(f^*).
\]

\( f^* \) is called the *Bayes decision rule*.

Notice that any choice for \( f^*(x) \) is equally good when \( \eta(x) = 1/2 \), so there can be several Bayes decision rules.
The excess risk of a decision rule (above the Bayes risk) can be quantified in terms of a certain distance from $f^*$.

**Theorem:** For any $f : \mathcal{X} \to \mathcal{Y}$,

$$R(f) - R(f^*) = E(1[f(X) \neq f^*(X)]|2\eta(X) - 1).$$
Risk and distance from $f^*$: Proof

We have seen $R(f) = E \left( 1[f(X) \neq 1](2\eta(X) - 1) + 1 - \eta(X) \right)$.

Hence,

$$R(f) - R(f^*) = E \left( 1[f(X) \neq 1] - 1[f^*(X) \neq 1] \right)(2\eta(X) - 1).$$

But

$$(1[f(X) \neq 1] - 1[f^*(X) \neq 1])(2\eta(X) - 1)$$

$$= 1[f(X) \neq f^*(X)](1[f(X) \neq 1] - 1[f^*(X) \neq 1])(2\eta(X) - 1)$$

$$= \begin{cases} 
1[f(X) \neq f^*(X)](2\eta(X) - 1) & \text{if } 2\eta(X) - 1 \geq 0, \\
1[f(X) \neq f^*(X)](-1)(2\eta(X) - 1) & \text{if } 2\eta(X) - 1 < 0.
\end{cases}$$

(from the definition of $f^*$)

$$= 1[f(X) \neq f^*(X)]|2\eta(X) - 1|,$$
Plug-in methods

This suggests one family of pattern classification methods: *plug-in* methods:

- Use the data to come up with an estimate \( \hat{\eta} \) of \( \eta \),
- Choose

\[
 f_{\hat{\eta}}(x) = \begin{cases} 
 1 & \text{if } \hat{\eta}(x) \geq 1/2, \\
 -1 & \text{otherwise}.
\end{cases}
\]
In estimating $\eta$, what criterion should we aim to minimize? $L_1(\mu)$ distance between $\hat{\eta}$ and $\eta$ suffices:

**Theorem:** For any $\hat{\eta} : \mathcal{X} \rightarrow \mathbb{R}$,

$$R(f_{\hat{\eta}}) - R^* \leq 2E |\eta(X) - \hat{\eta}(X)|.$$
Plug-in methods: Proof

We have seen:

\[ R(f_{\hat{\eta}}) - R^* = 2E \mathbb{1}[f_{\hat{\eta}}(X) \neq f^*(X)]|\eta(X) - 1/2|. \]

Now, if \( f_{\hat{\eta}}(X) \neq f^*(X) \), then \( \hat{\eta}(X) \) and \( \eta(X) \) must lie on opposite sides of 1/2, so

\[ |\eta(X) - \hat{\eta}(X)| = |\eta(X) - 1/2| + |\hat{\eta}(X) - 1/2| \geq |\eta(X) - 1/2|. \]

Thus, when \( f_{\hat{\eta}}(X) \neq f^*(X) \), we have

\[ 1[f_{\hat{\eta}}(X) \neq f^*(X)]|\eta(X) - 1/2| \leq |\eta(X) - \hat{\eta}(X)| \]

And this inequality is trivially true when the indicator is zero. Hence,

\[ R(f_{\hat{\eta}}) - R^* = 2E \mathbb{1}[f_{\hat{\eta}}(X) \neq f^*(X)]|\eta(X) - 1/2| \leq 2E|\eta(X) - \hat{\eta}(X)|. \]
Estimating $\eta$ is not necessary

Notice that estimating $\eta$ accurately is not necessary for accurate classification. In particular, this bound for a plug-in classifier can be very loose. For example, if $\eta(X) \in \{0, 1\}$, then for any $\epsilon > 0$, there is a $\hat{\eta}$ satisfying

- $\hat{\eta}$ and $\eta$ are always on the same side of $\frac{1}{2}$, and
- $|\hat{\eta}(X) - \eta(X)| = \frac{1-\epsilon}{2}$ a.s.

So

$$R(f_{\hat{\eta}}) - R^* = 0 \ll 1 - \epsilon = 2E|\eta(X) - \hat{\eta}(X)|.$$  

That is, the bound might be vacuous even though the classifier is optimal.
Choosing from a class of decision rules

An alternative to modelling the conditional distribution $\eta$ of $Y$ given $X$: fix a class $F$ of decision rules (functions from $\mathcal{X}$ to $\mathcal{Y}$) and use the data to choose $f_n$ from $F$.

For example, consider the class of linear threshold functions on $\mathcal{X} = \mathbb{R}^d$,

$$ F = \{ x \mapsto \text{sign}(\theta'x) : \theta \in \mathbb{R}^d \} . $$

The decision boundaries are hyperplanes through the origin ($d - 1$-dimensional subspaces), and the decision regions are half-spaces through the origin. (PICTURE)
Linear threshold functions

For thresholded \textit{linear} functions, the decision boundaries are hyperplanes through the origin. For thresholded \textit{affine} functions, the decision boundaries are arbitrary hyperplanes.

Essentially equivalent:

\[
F = \left\{ x \mapsto \text{sign}(\theta'x + c) : \theta \in \mathbb{R}^d, c \in \mathbb{R} \right\}
\]
\[
= \left\{ x \mapsto \text{sign}(\tilde{\theta}'\tilde{x}) : \tilde{\theta} \in \mathbb{R}^{d+1} \right\},
\]

where we define \( \tilde{x}' = (x'1) \). For notational simplicity, we’ll stick to the linear case.
Empirical risk minimization

How can we choose \( f \in F \)? One approach is *empirical risk minimization*:

Choose \( f \) from \( F \) to minimize the *empirical risk*,

\[
\hat{R}(f) = \hat{E}\ell(f(X), Y) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i).
\]
Consider empirical risk minimization over the class of linear threshold functions.

**Approximation** Very restricted class of decision rules. Can consider a much bigger class, and retain many of the attractive properties of linearly parameterized functions, by considering a nonlinear transformation \( \phi : \mathbb{R}^d \to \mathbb{R}^D \) for some \( D \gg d \). (Kernel methods.)

**Estimation** Small \( d/n \) is ok. Large can also be ok if we regularize.

**Computation** Easy if \( \hat{R}(f) = 0 \). In general, hard if not. Can simplify if we consider alternative (convex) loss functions \( \ell \).
Perceptron algorithm

Input: \((X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \{\pm 1\}\)

\(\theta_0 = 0 \in \mathbb{R}^d, t = 0\)

while some \((x_i, y_i)\) is misclassified, i.e., \(y_i \neq \text{sign}(\theta_t^T x_i)\)

pick some misclassified \((x_i, y_i)\)

\(\theta_{t+1} := \theta_t + y_i x_i\)

\(t := t + 1\)

Return \(\theta_t\).

Here,

\[
\text{sign}(\alpha) = \begin{cases} 
1 & \alpha > 0, \\
-1 & \alpha < 0, \\
0 & \alpha = 0.
\end{cases}
\]
Perceptron convergence theorem

**Theorem:** Given *linearly separable data* (i.e., there is a $\theta \in \mathbb{R}^d$ such that for all $i$, $y_i \theta^T x_i > 0$), for any choices made at the update step, it terminates (with empirical risk zero) after no more than $\frac{R^2}{\gamma^2}$ updates, where

$$R = \max_i \|x_i\|, \quad \text{(radius of data)}$$

$$\gamma = \min_i \frac{\theta^T x_i y_i}{\|\theta\|}, \quad \text{(margin)}$$
The idea is to use the inner product $\theta_t^T \theta$ as a measure of progress, and show that each mistake gives a big increase to the inner product (aligns $\theta_t$ with $\theta$), but gives only a small increase to $\|\theta_t\|$.

First,

$$\theta_{t+1}^T \theta = (\theta_t + y_i x_i)^T \theta$$

$$\geq \theta_t^T \theta + \gamma \|\theta\|.$$  

But $\theta_0 = 0$, so $\theta_t^T \theta \geq t \gamma \|\theta\|$.
On the other hand,

\[ \| \theta_{t+1} \|^2 = \| \theta_t + y_i x_i \|^2 \]
\[ = \| \theta_t \|^2 + \| x_i \|^2 + 2 y_i \theta_t^T x_i \]
\[ \leq \| \theta_t \|^2 + R^2. \]

But \( \theta_0 = 0 \), so \( \| \theta_t \|^2 \leq tR^2 \).

Combining (and using Cauchy-Shwarz):

\[ t \gamma \| \theta \| \leq \theta_t^T \theta \leq \| \theta_t \| \| \theta \| \leq \sqrt{tR} \| \theta \|. \]
Linear threshold functions

For linearly separable data (i.e., there is a $\theta \in \mathbb{R}^d$ such that for all $i$, $y_i \theta^T x_i > 0$), finding an empirical risk minimizer corresponds to finding a point satisfying $n$ linear inequalities:

$$y_i \theta^T x_i > 0.$$ 

In particular, it can be solved with a linear program:

$$\max_{\gamma, \theta} \quad \gamma$$

subject to

$$y_i \theta^T x_i \geq \gamma.$$ 

So we can find a solution in polynomial time (even though the optimal $\gamma$ might be exponentially small, so the perceptron algorithm might take exponential time).
Overview

1. Pattern classification: $\mathcal{Y} = \{\pm 1\}$.

2. Plug-in estimators: $R(f_{\hat{\eta}}) - R^* \leq 2E |\eta(X) - \hat{\eta}(X)|$.


4. Linear threshold functions.

5. Perceptron algorithm: convergence.