Handout 8
Binary Classification. Policy Learning

Instructor: Vira Semenova Note author: Vira Semenova

1 Binary Classification

I introduce binary classification problem. Consider a sequence $(X_i, Y_i)_{i=1}^n$ of n i.i.d draws from a joint distribution $P_{X,Y}$. Here, X is a covariate vector and $Y \in \{1, 0\}$ is a binary outcome where $Y \mid X \sim B(p(X))$, where

$$p(X) = \Pr(Y = 1 \mid X).$$

In what follows, we adopt zero-one loss function to evaluate classification mistakes. Specifically, we incur loss 1 if $Y \neq h(X)$ (i.e., the actual label is different from predicted one) and 0 otherwise (if they coincide). Aggregating over $P_{X,Y}$ gives ex-ante risk

$$R(h) = \Pr(Y \neq h(X)).$$

The goal is to find an optimal classifier

$$h^* = \arg \min_{\text{all classifiers}} R(h) \tag{1.1}$$

Definition 1 (Bayes classifier)**.** *The Bayes classifier of X given Y, denoted h^*, is the function defined by the rule*

$$h^*(x) = \begin{cases} 1 \ \textit{if}\ p(x) > 1/2 \\ 0 \ \textit{if}\ p(x) \leq 1/2 \end{cases} \tag{1.2}$$

In other words, $h^*(X) = 1$ if $\Pr(Y = 1 \mid X) > \Pr(Y = 0 \mid X)$ and zero otherwise.
The Bayes risk – the risk of Bayes classifier h^* – is

$$\begin{aligned} R(h^*) &= \mathbb{E}\left[\Pr(Y = 1 \mid X : p(X) \leq 1/2) + \Pr(Y = 0 \mid X : p(X) \geq 1/2)\right] \\ &= \mathbb{E}\left[p(X)1\{p(X) \leq 1/2\} + (1 - p(X))1\{p(X) \geq 1/2\}\right] \\ &= \mathbb{E}\left[\min(p(X), 1 - p(X))\right] \leq 1/2. \end{aligned}$$

Theorem 1. *The Bayes risk is the smallest admissible risk among all classifiers. In other words, the classifier $h^*(X)$ is the first-best (optimal) classifier. In particular, for any classifier h,*

$$\mathcal{E}(h) := R(h) - R(h^*) = \mathbb{E}\left[|2p(X) - 1| 1\{h(X) \neq h^*(X)\}\right] \geq 0. \tag{1.3}$$

Proof. The proof is left as an exercise. Try proving (1.3) in 3 steps:

1. Show that the Bayes risk is $R(h^*) = \mathbb{E}\left[\min(p(X), 1 - p(X))\right]$
2. Show that the excess risk of a classifier h is

$$R(h) - R(h^*) = \mathbb{E}\left[(2p(X) - 1)(1\{h(X) = 0\} - 1\{h^*(X) = 0\})\right].$$

3. Conclude that (1.3) holds.

For the full proof, see [?], Chapter 2. □

We make several remarks. First, the quantity $\mathcal{E}(h) := R(h) - R(h^*)$ in the statement of the theorem above is called the *excess risk* of h above the Bayes risk. The theorem implies that $R(h) - R(h^*) \geq 0$. Second, the risk of the Bayes classifier $R(h^*)$ equals $1/2$ if and only if $p(X) = 1/2$ almost surely. This maximal risk for the Bayes classifier occurs precisely when X "contains no information" about the outcome Y. Equation (1.3) makes clear that the excess risk weighs the discrepancy between h and h^* according to how far h is from $1/2$. When $p(X)$ is close to $1/2$, no classifier can perform well and the excess risk is low. When $p(X)$ is far from $1/2$, the Bayes classifier performs well and we penalize classifiers that fail to do so more heavily.

1.1 Empirical Risk Minimization

The Bayes classifier h^*, while optimal, presents a major drawback: we cannot compute it because we do not know the regression function $p(x)$. Instead, we have access to the data $(X_i, Y_i)_{i=1}^n$, which contains some (but not all) information about p and thus h^*. In order to mimic the properties of h^*. recall that it minimizes $R(h)$ over all h. A generative/estimation approach would be to estimate the conditional probability $p(X)$ and plug it into (1.2). In the language of ML, this is called building a generative model on $p(X)$. Such a model would require various assumptions on $p(X)$, such as smoothness or sparsity. Here, we consider an alternative – discriminative (machine learning) approach – one makes assumptions on what classifiers are likely to perform correctly.

Definition 2 (Empirical Risk Minimization). *Given the data $(W_i)_{i=1}^n = (X_i, Y_i)_{i=1}^n$ define the empirical risk*

$$\widehat{R}_n(h) := n^{-1} \sum_{i=1}^n 1\{h(X_i) \neq Y_i\}.$$

An Empirical Risk Minimization (ERM) classifier is

$$\widehat{h}_n := \arg\min_{\mathcal{H}} \widehat{R}_n(h), \tag{1.4}$$

where $\mathcal{H}$ is a set of classifier to search for.

Assume that we are given a class $\mathcal{H}$ in which we expect to find a classifier that performs well. This class may be constructed from domain knowledge or simply computational convenience. Ë We will see some examples in the class. For any candidate classifier $\widehat{h}_n$ built from the data, we can decompose its excess risk as follows:

$$\mathcal{E}(\widehat{h}_n) = R(\widehat{h}_n) - R(h^*) = \underbrace{R(\widehat{h}_n) - \inf_{h\in\mathcal{H}} R(h)}_{\text{estimation error}} + \underbrace{\inf_{h\in\mathcal{H}} R(h) - R(h^*)}_{\text{approximation error}}.$$

On the one hand, estimation error accounts for the fact that we only have a finite amount of observations and thus a partial knowledge of the distribution $P_{X,Y}$. Hopefully we can drive this error to zero as $n \to \infty$. But we already know from the no-free-lunch theorem that this will not happen if $\mathcal{H}$ is the set of all classifiers. Therefore, we need to take $\mathcal{H}$ small enough. On the other hand, if$\mathcal{H}$ is too small, it is unlikely that we will find classifier with performance close to that of h^*. A tradeoff between estimation and approximation can be made by letting $\mathcal{H} = \mathcal{H}_n$ grow (but not too fast) with n.

For now, suppose $\mathcal{H}$ is fixed and $h^* \in \mathcal{H}$. Decompose

$$\begin{aligned}\mathcal{E}(\widehat{h}_n) &= \left(\widehat{R}_n(\widehat{h}_n) - \widehat{R}(h^*)\right) + \left(R(\widehat{h}_n) - \widehat{R}_n(\widehat{h}_n)\right) + \left(\widehat{R}_n(h^*) - R(h^*)\right) \\ &\leq (\leq 0) + |\widehat{R}_n(\widehat{h}_n) - R(\widehat{h}_n)| + \underbrace{|\widehat{R}_n(h^*) - R(h^*)|}_{\Rightarrow^p 0 \text{ by LLN}}.\end{aligned}$$

Note that $\widehat{h}_n$ is a random quantity, and $\widehat{R}_n(\widehat{h}_n)$ is **NOT** a sample average of i.i.d random draws. The middle term *requires uniform* convergence

$$|\widehat{R}_n(\widehat{h}_n) - R(\widehat{h}_n)| \leq \sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)|,$$

where the rate of convergence depends on the complexity of $\mathcal{H}$. If the class $\mathcal{H}$ has at most $|\mathcal{H}| = M$ elements,

2 Statistical treatment rules. Policy Learning

In this section, I will consider a notable application of ERM to economics: statistical treatment rules and welfare maximization. [1]. . An important objective of empirical analysis of experimental and quasi-experimental data is to determine the individuals who should be treated based on their observable characteristics. A statistical treatment rule $h(X) \to \{1, 0\}$ is to assign the decision rule "to treat" if $\delta(X) = 1$ and "do not treat" if $h(X) = 0$. Since $h(X)$ is binary, it is convenient to talk about *decision sets* $G \subset \mathcal{X}$ such that

$$h(X) = 1 \text{ if and only if } X \in G.$$

The subject's potential outcome when treated is $Y(1)$ and when not treated is $Y(0)$. The realized outcome Y is

$$Y = DY(1) + (1 - D)Y(0).$$

Define the propensity score (i.e., probability of treatment assignment as)

$$\mu_1(X) = \Pr(D = 1 \mid X), \quad \mu_0(X) = 1 - \mu_1(X). \tag{2.1}$$

Define the conditional average treatment effect as

$$\tau(X) := \mathbb{E}\left[Y \mid D = 1, X\right] - \mathbb{E}\left[Y \mid D = 0, X\right].$$

The average welfare of a classifier G is

$$\begin{aligned} W(G) &= \mathbb{E}\left[Y(1)1\{X \in G\} + Y(0)1\{X \in G^c\}\right] \\ &= \mathbb{E}\left[\frac{DY}{\mu_1(X)}1\{X \in G\} + \frac{(1-D)Y}{\mu_0(X)}1\{X \in G^c\}\right] \end{aligned}$$

Theorem 2 (First-best Decision Rule)**.** *The first-best (optimal) decision rule*

$$\begin{aligned} G^* &= \arg\min_{G \in \mathcal{G}} W(G) \\ &= \{X : \tau_0(X) \geq 0\}. \end{aligned}$$

Proof.

$$\begin{aligned} W(G) &= \mathbb{E}\left[\frac{DY}{\mu_1(X)}1\{X \in G\} + \frac{1-D}{\mu_0(X)}1\{X \in G^c\}\right] \\ &= \mathbb{E}\left[\left(\frac{DY}{\mu_1(X)} - \frac{1-D}{\mu_0(X)}\right)1\{X \in G\}\right] + \mathbb{E}\left[\frac{(1-D)Y}{\mu_0(X)}\right] \\ &= \mathbb{E}\left[\tau_0(X)1\{X \in G\}\right] + \mathbb{E}\left[\frac{(1-D)Y}{\mu_0(X)}\right]. \end{aligned}$$

Therefore, G^* maximizes $W(G)$ if $X \in G$ if and only if $\tau_0(X) \geq 0$. □

Theorem 2 establishes the optimality of first-best decision rule. The quantity

$$R(G) = W(G) - W(G^*)$$

is the *regret* of decision rule G relative to the optimal rule G^*.

Next, I define the EWM rule analogous to the ERM in statistical learning. The data $(W_i)_{i=1}^n = (X_i, D_i, Y_i)_{i=1}^n$ where

$$Y_i = \begin{cases} Y_{1,i} & D_i = 1 \\ Y_{0,i} & D_i = 0 \end{cases}$$

In particular, only one outcome of the two is observed. This setting is called a *partial feedback*. Nevertheless, it is possible to construct a decision rule

[1] This section is based on Section 2 of [Kitagawa and Tetenov, 2018]

Definition 3 (Empirical Welfare Maximization, [Kitagawa and Tetenov, 2018])**.** *Given the data $(W_i)_{i=1}^n = (X_i, Y_i)_{i=1}^n$ define the empirical risk*

$$\widehat{W}_n(G) := n^{-1} \sum_{i=1}^{n} \frac{D_i Y_i}{\mu_1(X_i)} 1\{X_i \in G\} + \frac{1 - D_i}{\mu_0(X_i)} 1\{X_i \in G^c\}$$

An Empirical Risk Minimization (ERM) classifier is

$$\widehat{G} := \arg\min_{\mathcal{G}} \widehat{W}_n(G) \tag{2.2}$$

where $\mathcal{G}$ is a set of classifier to search for.

Definition 4 (Penalized Welfare Maximization, [Mbakop and Tabord-Meehan, 2021])**.** *An Penalized Empirical Risk Minimization (ERM) classifier is*

$$\widehat{G} := \arg\min_{\mathcal{G}} \widehat{W}_n(G) + C_n(k), \tag{2.3}$$

where $\mathcal{G}$ is a set of classifier to search for and $C_n(k)$ is complexity penalty.

References

[Kitagawa and Tetenov, 2018] Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86:591–616.

[Mbakop and Tabord-Meehan, 2021] Mbakop, E. and Tabord-Meehan, M. (2021). Model selection for treatment choice: Penalized welfare maximization. *Econometrica*, 89:825–848.