13 Classification

Classification is the problem of predicting a binary variable $Y \in \{0, 1\}$ from features X (or a discrete outcome $Y \in \{1, ..., K\}$ in multi-class classification). It is also sometimes called discrimination, pattern classification, or pattern recognition. A classification rule or classifier is a function $h : \mathcal{X} \to \{0, 1\}$ that predicts the outcome from features, and the typical choice of error is misclassification error

$$R(h) = \mathbb{P}\Big\{Y \neq h(X)\Big\}$$

i.e., the probability of guessing the wrong label.

1 The optimal classifier

Our first result characterizes the optimal classifier, which is called the Bayes classifier. Many of the results of this and following sections can be found in Devroye et al. [1996].

Theorem 1. *Let* $h^*(x) = \mathbb{1}\{\mu(x) > 1/2\}$ *for* $\mu(x) = \mathbb{E}(Y \mid X = x)$. *Then*

$$\mathbb{P}\Big\{Y \neq h^*(X)\Big\} \leq \mathbb{P}\Big\{Y \neq h(X)\Big\}$$

for any function $h : \mathcal{X} \to \{0, 1\}$.

Proof. We have for any h that

$$\mathbb{P}\{Y \neq h(X) \mid X\} = \mu(X)\{1 - h(X)\} + \{1 - \mu(X)\}h(X)$$

so that

$$\begin{aligned}
&\mathbb{P}\{Y \neq h(X) \mid X = x\} - \mathbb{P}\{Y \neq h^*(X) \mid X = x\} \\
&\qquad = \Big\{h^*(x) - h(x)\Big\}\mu(x) + \Big\{h(x) - h^*(x)\Big\}\{1 - \mu(x)\} \\
&\qquad = \Big\{h^*(x) - h(x)\Big\}\Big\{2\mu(x) - 1\Big\} \\
&\qquad = \begin{cases} \{1 - h(x)\}\{2\mu(x) - 1\} & \text{if } 2\mu(x) > 1 \\ -h(x)\{2\mu(x) - 1\} & \text{if } 2\mu(x) \leq 1 \end{cases} \\
&\qquad \geq 0
\end{aligned}$$

so the result follows after taking an expectation over X. □

Note if X has a density then, letting $p_y(x) = p(x \mid Y = y)$ and $\pi_y = \mathbb{P}(Y = y)$, we have

$$\mu(x) = \frac{\pi_1 p_1(x)}{\pi_1 p_1(x) + \pi_0 p_0(x)}$$

and the Bayes classifier can also be written as

$$h^*(x) = \mathbb{1}\left\{\frac{p_1(x)}{p_0(x)} > \frac{\pi_0}{\pi_1}\right\}.$$

Proposition 1. *The risk of the optimal classifier can be expressed as*

$$\begin{aligned} R(h^*) &= \inf_{h:\mathcal{X}\to\{0,1\}} \mathbb{P}\Big\{Y \neq h(X)\Big\} \\ &= \mathbb{E}\left[\min\Big\{\mu(X), 1-\mu(X)\Big\}\right] \\ &= 1/2 - \mathbb{E}|2\mu(X)-1|/2 \end{aligned}$$

and if X has a density

$$R(h^*) = \int \min\Big\{\pi_1 p_1(x), \pi_0 p_0(x)\Big\}\, dx.$$

Proof. We have

$$\begin{aligned} \mathbb{E}\left[\min\Big\{\mu(X), 1-\mu(X)\Big\}\right] &= \mathbb{E}\Big[\mu(X)\mathbb{1}\{\mu(X) \leq 1-\mu(X)\} + \{1-\mu(X)\}\mathbb{1}\{1-\mu(X) < \mu(X)\}\Big] \\ &= \mathbb{E}\Big[\mu(X)\{1-h^*(X)\} + \{1-\mu(X)\}h^*(X)\Big] \\ &= \mathbb{E}\Big[Y\{1-h^*(X)\} + (1-Y)h^*(X)\Big] = \mathbb{P}\{Y \neq h^*(X)\} \end{aligned}$$

$$\begin{aligned} 1/2 - \mathbb{E}|2\mu(X)-1|/2 &= 1/2 - \mathbb{E}\Big[\{2\mu(X)-1\}h^*(X) - \{2\mu(X)-1\}\{1-h^*(X)\}\Big]/2 \\ &= 1/2 - \mathbb{E}\Big[\{2\mu(X)-1\}\{2h^*(X)-1\}\Big]/2 = 1/2 - \mathbb{E}\Big[(2Y-1)\{2h^*(X)-1\}\Big]/2 \\ &= \mathbb{E}\Big[\mathbb{1}\{Y \neq h^*(X)\}\Big] = \mathbb{P}\{Y \neq h^*(X)\} \end{aligned}$$

and the last since $\mu(x) = \mathbb{P}(Y = y \mid X = x) = p(x \mid Y = y)\mathbb{P}(Y = y)/p(x)$. □

2 Plug-in estimation

Generic regression estimators can be turned into classifiers, based on the form of the optimal classifier, via the plug-in estimator

$$\widehat{h}(X) = \mathbb{1}\Big\{\widehat{\mu}(X) > 1/2\Big\}.$$

The error of this classifier can be related to that of the regression estimator itself, as in the following result, which shows that in some sense classification is not more difficult than regression.

Theorem 2. *Let $\widehat{h}$ and h^* denote the plug-in estimator and optimal Bayes classifier, respectively. Then*

$$R(\widehat{h}) - R(h^*) \leq 2\|\widehat{\mu} - \mu\|_1 \leq 2\|\widehat{\mu} - \mu\|_2$$

for $\|f\|_1 = \int |f(x)| d\mathbb{P}(x)$ and $\|f\|_2 = \sqrt{\int f(x)^2\, d\mathbb{P}(x)}$ the $L_1(\mathbb{P})$ and $L_2(\mathbb{P})$ norms, respectively.

Proof. We have

$$\begin{aligned}
&\mathbb{P}\{Y \neq \widehat{h}(X) \mid X\} - \mathbb{P}\{Y \neq h^*(X) \mid X\} \\
&\qquad = \mathbb{E}\Big[Y\{h^*(X) - \widehat{h}(X)\} + (1-Y)\{\widehat{h}(X) - h^*(X)\} \mid X\Big] \\
&\qquad = |2\mu(X) - 1| \mathbb{1}\{h^*(X) \neq \widehat{h}(X)\}
\end{aligned}$$

so that

$$\begin{aligned}
R(\widehat{h}) - R(h^*) &= \int |2\mu(x) - 1| \mathbb{1}\{h^*(x) \neq \widehat{h}(x)\} \; d\mathbb{P}(x) \\
&\leq \int 2|\widehat{\mu}(x) - \mu(x)| \mathbb{1}\{h^*(x) \neq \widehat{h}(x)\} \; d\mathbb{P}(x) \\
&\leq 2 \int |\widehat{\mu}(x) - \mu(x)| \; d\mathbb{P}(x).
\end{aligned}$$

where in the second line we used the fact that $|\mu(x) - 1/2| \leq |\widehat{\mu}(x) - \mu(x)|$ when $h^* \neq \widehat{h}$ (since in that case $2\mu - 1$ and $2\widehat{\mu} - 1$ have opposite sign, which implies $|2\mu - 1| + |2\widehat{\mu} - 1| = |2(\widehat{\mu} - \mu)|$). The L_2 upper bound follows since $\mathbb{E}(Z^2) \geq \mathbb{E}(Z)^2$. □

A similar result can be obtained with the density-based plug-in estimator

$$\widehat{h}(x) = \mathbb{1}\left\{\frac{\widehat{p}_1(x)}{\widehat{p}_0(x)} > \frac{\widehat{\pi}_0}{\widehat{\pi}_1}\right\}.$$

Theorem 3. *Let $\widehat{h}$ and h^* denote the density-based plug-in estimator and optimal Bayes classifier, respectively. Then*

$$R(\widehat{h}) - R(h^*) \leq 2 \sum_{y=0}^{1} \int \Big|\widehat{\pi}_y \widehat{p}_y(x) - \pi_y p_y(x)\Big| dx.$$

Proof. The density-based plug-in estimator is equivalent to using the regression plug-in with

$$\widehat{\mu}(x) = \frac{\widehat{\pi}_1 \widehat{p}_1(x)}{\widehat{\pi}_1 \widehat{p}_1(x) + \widehat{\pi}_0 \widehat{p}_0(x)}$$

so, letting $f_y = \pi_y p_y(x)$ and $\widehat{f_y}$ the estimated version, we can apply the previous theorem to obtain

$$\begin{aligned}
R(\widehat{h}) - R(h^*) &\leq 2 \int \left| \frac{\widehat{f}_1}{\widehat{f}_1 + \widehat{f}_0} - \frac{f_1}{f_1 + f_0} \right| (f_1 + f_0) \; dx \\
&= 2 \int \left| \frac{\widehat{f}_1 (f_1 + f_0)}{\widehat{f}_1 + \widehat{f}_0} - f_1 \right| \; dx = 2 \int \left| \frac{\widehat{f}_1 f_0 - f_1 \widehat{f}_0}{\widehat{f}_1 + \widehat{f}_0} \right| \; dx \\
&= 2 \int \left| \frac{\widehat{f}_1 (f_0 - \widehat{f}_0) - (f_1 - \widehat{f}_1) \widehat{f}_0}{\widehat{f}_1 + \widehat{f}_0} \right| \; dx \\
&\leq 2 \int |\widehat{f}_1 - f_1| \; dx + 2 \int |\widehat{f}_0 - f_0| \; dx
\end{aligned}$$

□

Note the misclassification error $R(\widehat{h})$ is random, as it is a function of $\widehat{h}$; for a deterministic risk measure it is common to work with the excess risk

$$\mathcal{E}(\widehat{h}) = \mathbb{E}R(\widehat{h}) - R(h^*)$$

i.e., the expectation of the misclassification error relative to that of the Bayes classifier.

The two results above show that one can obtain, for free, regression and density estimation rates in classification. For example if we assume the regression function $\mu(x)$ or the class conditional densities $p_y(x)$ are s-smooth, then we can for example obtain

$$\mathcal{E}(\widehat{h}) = \mathbb{E}R(\widehat{h}) - R(h^*) \leq Cn^{-1/(2+d/s)}$$

with local polynomial or series estimators, and appropriate tuning parameter selection. In fact Yang [1999] shows these rates are already minimax optimal, i.e., that

$$\inf_{\widehat{h}} \sup_{P \in \mathcal{P}} \mathcal{E}_P(\widehat{h}) \gtrsim n^{-1/(2+d/s)}$$

for $\mathcal{P}$ the set of distributions with s-smooth regression functions $\mu(x)$ or class-conditional densities $p_y(x)$, and so these rates cannot be improved in a minimax sense, without adding more structure to the model.

3 Empirical risk minimization

An alternative approach for constructing estimators is to use empirical risk minimization (ERM). Specifically, we first posit a class of possible decision sets $\mathcal{G}$, which could be for example separating hyperplanes

$$\mathcal{G}_{lin} = \{x : \beta^{\mathrm{T}} x > 0, \beta \in \mathbb{R}^p\}$$

or trees, etc. Then we estimate the misclassification error across all sets, and find the set giving the best estimate, i.e.,

$$\widehat{G} = \arg\min_{G \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\Big[Y_i \neq \mathbb{1}\{X_i \in G\}\Big] = \arg\min_{G \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \Big[Y_i - \mathbb{1}\{X_i \in G\}\Big]^2.$$

Then the final classifier is given by

$$\widehat{h}(x) = \mathbb{1}(x \in \widehat{G}).$$

The following Theorems 4.5 and 4.6 from Devroye et al. [1996] give error rates for linear classifiers in particular.

Theorem 4. *Let $\widehat{G} = \arg\min_{G \in \mathcal{G}_{lin}} \widehat{R}(G)$ and $G^* = \arg\min_{G \in \mathcal{G}_{lin}} R(G)$. If $n \geq d$ and X has a density then*

$$\mathbb{E}\{R(\widehat{G}) - R(G^*)\} \leq \sqrt{\frac{2}{n}\Big\{(d+1)\log n + (2d+2)\Big\}}.$$

If the distribution is additionally linearly separable, so that $R(G^*) = 0$, then

$$\mathbb{E}\{R(\widehat{G}) - R(G^*)\} = \mathbb{E}\{R(\widehat{G})\} \leq \frac{d \log n + 2}{n - d}.$$

It is natural to consider what happens when we target the best decision set

$$G^* = \{x : \mu(x) > 1/2\}$$

as opposed to the best separating hyperplane $\arg\min_{G \in \mathcal{G}_{lin}} R(G)$, and/or we do not assume the distribution is linearly separable. The following result comes from Tsybakov [2004].

Remark 1. The following result imposes restrictions on the complexity of the decision set $G^* = \{x : \mu(x) > 1/2\}$, via metric entropy bounds. Recall the covering number $\mathcal{N}(\epsilon, \mathcal{F}, d)$ is the minimal number of d-distance (e.g., L_2) balls of radius ϵ needed to cover a set $\mathcal{F}$, and the metric entropy $\mathcal{H}(\epsilon, \mathcal{F}, d)$ is the logarithm of the covering number. VC-classes have ϵ-entropy logarithmic in $1/\epsilon$, and so are a relatively small class; smoothness classes have entropy $\epsilon^{-\rho}$ with $\rho = d/s$, for s the smoothness index (essentially the number of bounded derivatives).

Theorem 5. *Assume the decision set $G^* = \{x : \mu(x) > 1/2\}$ belongs to a class $\mathcal{G}$ of subsets of $\mathbb{R}^d$ whose metric entropy satisfies*

$$\mathcal{H}(\epsilon, \mathcal{G}, d_\Delta) \leq C\epsilon^{-\rho}, \ \forall \epsilon > 0$$

for some constant $0 < \rho < 1$, where $d_\Delta(A, B)$ is the Lebesgue measure of the set difference $A \Delta B = (A \setminus B) \cup (B \setminus A)$. Then for the ERM classifier $\widehat{G} = \arg\min_{G \in \mathcal{G}} \widehat{R}(G)$ we have

$$\mathbb{E}\Big\{R(\widehat{G}) - R(G^*)\Big\} = O\left(\frac{1}{\sqrt{n}}\right).$$

Remark 2. Note the assumption that $\rho < 1$ in the above result requires that $s > d$ in Holder-type classes, which can be a strong assumption. When $s > d$ in a regression estimation context, one can estimate μ at faster than $n^{-1/3}$ rates.

Remark 3. In some cases it is not feasible to compute an empirical risk minimizer over the whole class $\mathcal{G}$, because the class is too large. In that case one can minimize over a sufficiently fine ϵ-net, which is another class $\mathcal{G}'$ such that for any $G \in \mathcal{G}$ there is exists a $G' \in \mathcal{G}'$ such that $d(G, G') \leq \epsilon$ for some distance d.

Note the above results for ERM yield faster $1/\sqrt{n}$ or $1/n$ rates compared to the slower $n^{-1/(2+d/s)}$ rates we saw for the plug-in estimator under smoothness assumptions. This suggests that perhaps $1/n$ rates are optimal and that ERM estimators are somehow better than plug-in estimators, which may be suboptimal/deficient in some way. However, the above ERM results impose structure on the decision set G^*, while the plug-in results impose structure on the regression function $\mu(x)$, so the results are not directly comparable.

Further, as mentioned earlier, Yang [1999] (and Audibert and Tsybakov [2007]) showed that the plug-in rates are in fact optimal when only assuming the regression function is smooth. This at least shows that plug-in estimators are not deficient, and already optimal for classification, when only assuming smoothness of μ. Audibert and Tsybakov [2007] also show that plug-in-type estimators can achieve $1/\sqrt{n}$, $1/n$, or even exponentially fast rates under smoothness, when we add a margin assumption, which will be discussed in later sections.

4 Cross-validation

Similar results regarding cross-validated model selection in regression can be obtained in classification contexts as well. For simplicity we discuss the sample splitting approach.

We split the sample into two parts,

$$D_1^n \quad \text{and} \quad D_2^n$$

of size n_1 and n_2, respectively, using D_1^n for training and using D_2^n as a test sample for estimating a best classifier. Specifically, we have a set $\mathcal{H}_n = \{h_1, h_2, ..., \}$ of classifiers built from the training sample D_1^n. Then moving to the test data, we pick the best classifier satisfying, for example,

$$\widehat{h} = \underset{h \in \mathcal{H}_n}{\arg\min} \frac{1}{n_2} \sum_{i \in D_2^n} \mathbb{1}\Big\{Y_i \neq h(X_i)\Big\} = \underset{h \in \mathcal{H}_n}{\arg\min} \frac{1}{n_2} \sum_{i \in D_2^n} \{Y_i - h(X_i)\}^2.$$

A fundamental question is how the risk of this classifier compares to that of an oracle who had used the best (random) choice

$$\widetilde{h} = \underset{h \in \mathcal{H}}{\arg\min}\, R(h).$$

Note this $\widetilde{h}$ is called the best random choice since it is a random variable through its dependence on D_1^n, via the sequence of estimators in $\mathcal{H}_n$. However, also note that it is not itself a feasible choice in practice since it depends on the unknown distribution $\mathbb{P}$.

Theorem 6. *For $\widehat{h}$ and $\widetilde{h}$ defined above, we have*

$$\mathbb{E}\{R(\widehat{h}) - R(h^*)\} \leq \mathbb{E}\{R(\widetilde{h}) - R(h^*)\} + \sqrt{\frac{8(1 + \log 2|\mathcal{H}_n|)}{n_2}}.$$

Proof. As in the regression setting introduce the quantity

$$\begin{aligned} Z(h) = Z(h; D_1^n) &\equiv \Big\{Y - h(X)\Big\}^2 - \Big\{Y - h^*(X)\Big\}^2 \\ &= \Big\{h^*(X) - h(X)\Big\}\Big\{Y - h^*(X) + Y - h(X)\Big\} \in \pm 2 \end{aligned}$$

Using the same decomposition as in the regression case, we have

$$R_n = P\{Z(\widehat{h})\} = \mathbb{P}_n\{Z(\widehat{h})\} + (P - \mathbb{P}_n)Z(\widehat{h}) \equiv S_n + T_n$$

and by the same logic as before,

$$S_n = \mathbb{P}_n\{Z(\widehat{h})\} \leq \mathbb{P}_n\{Z(\widetilde{h})\}$$

so that

$$\mathbb{E}(S_n \mid D_1^n) \leq \mathbb{E}\Big[\mathbb{P}_n\{Z(\widetilde{h})\} \;\Big|\; D_1^n\Big] = R(\widetilde{h}) - R(h^*).$$

Via a union bound and Hoeffding's inequality we have

$$\begin{aligned}\mathbb{P}\Big\{|(P-\mathbb{P}_n)Z(\widehat{h})| \geq t \mid D_1^n\Big\} &\leq |\mathcal{H}_n| \max_{h\in\mathcal{H}_n} \mathbb{P}\Big\{|(P-\mathbb{P}_n)Z(h)| \geq t \mid D_1^n\Big\} \\ &\leq |\mathcal{H}_n| 2\exp\left\{\frac{-n_2 t^2}{(4)^2/2}\right\}.\end{aligned}$$

Now

$$\begin{aligned}\mathbb{E}(T_n \mid D_1^n) &\leq \sqrt{\mathbb{E}(T_n^2 \mid D_1^n)} \leq \sqrt{u + \int_u^\infty \mathbb{P}(T_n^2 \geq t \mid D_1^n)\, dt} \\ &= \sqrt{u + \int_u^\infty \mathbb{P}(|T_n| \geq \sqrt{t} \mid D_1^n)\, dt} \leq \sqrt{u + \int_u^\infty |\mathcal{H}_n| 2\exp\left(\frac{-n_2 t}{8}\right)\, dt} \\ &= \sqrt{u + 2|\mathcal{H}_n|\left\{\frac{-8}{n_2}\exp\left(\frac{-n_2 t}{8}\right)\right\}\Big|_u^\infty} = \sqrt{u + \frac{2(8)|\mathcal{H}_n|}{n_2}\exp\left(\frac{-n_2 u}{8}\right)}\end{aligned}$$

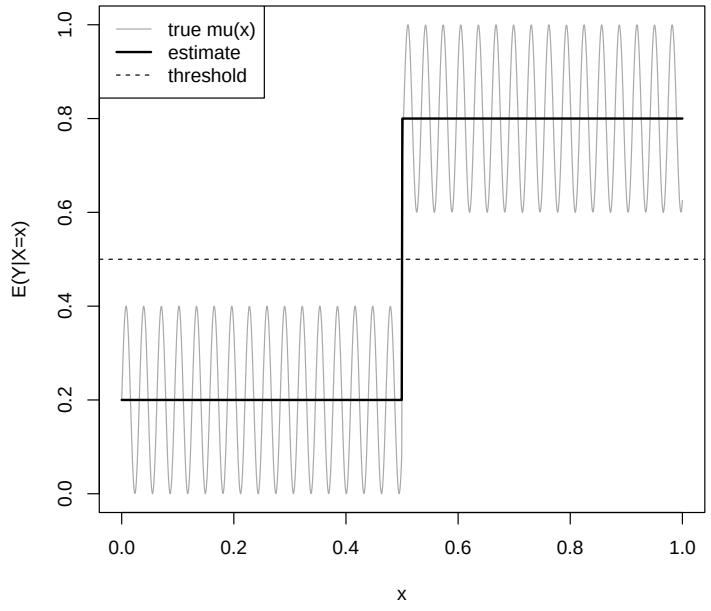
To minimize over u we take $u = 8\log 2|\mathcal{H}_n|/n_2$, which gives

$$\mathbb{E}(T_n \mid D_1^n) \leq \sqrt{\frac{8(1+\log 2|\mathcal{H}_n|)}{n_2}}.$$

and so yields the result. □

5 Margin assumption

Intuitively, classification should be easier than regression in some settings. Consider the following toy example, where the regression function is difficult to estimate accurately, but a simple inaccurate estimator leads to perfect classification, matching the oracle Bayes classifier.



Part of what makes this toy example easy with respect to classification is that the regression function is almost never close to $1/2$, which is where it is hardest/impossible to guess labels accurately. However, our previous analyses had no way of exploiting this kind of structure when present. The following margin assumption is one way of incorporating such structure; it has been used by Mammen and Tsybakov [1999], Tsybakov [2004], Audibert and Tsybakov [2007], etc. The latter paper provides the foundation for this section and the next.

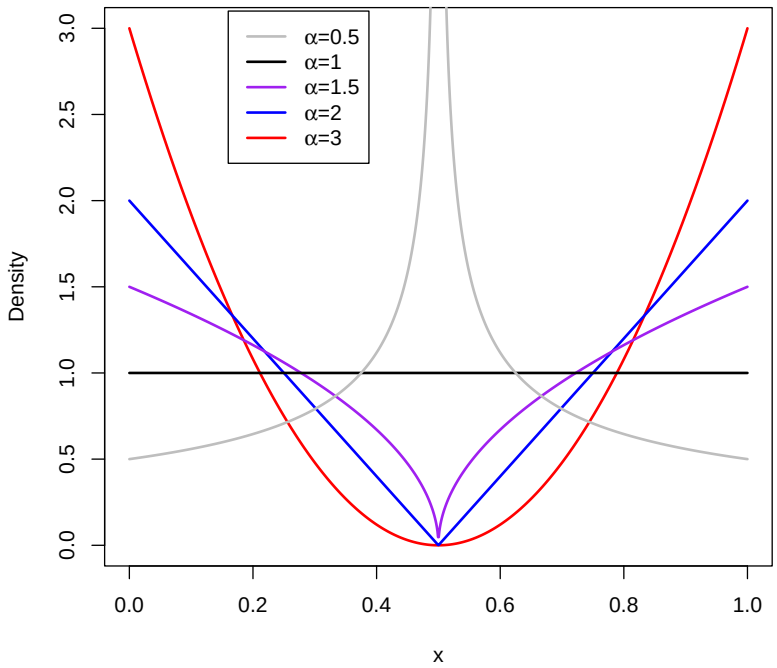
Assumption 1. There exist constants $C > 0$ and $\alpha \geq 0$ such that

$$\mathbb{P}(0 < |\mu(X) - 1/2| \leq t) \leq Ct^{\alpha}$$

for all $0 < t \leq t_0 \leq 1/2$.

The margin assumption controls how much mass of the distribution of $\mu(X)$ is near the problematic point $1/2$. The plot below shows some densities with different margins:

1. $f(x) = \frac{1}{2\sqrt{2}\sqrt{|x-1/2|}} \implies \mathbb{P}(0 < |X - 1/2| \leq t) = \sqrt{2}t^{1/2}$
2. $f(x) = 1 \implies \mathbb{P}(0 < |X - 1/2| \leq t) = 2t$
3. $f(x) = (3/\sqrt{2})\sqrt{|x - 1/2|} \implies \mathbb{P}(0 < |X - 1/2| \leq t) = (2\sqrt{2})t^{3/2}$
4. $f(x) = 4|x - 1/2| \implies \mathbb{P}(0 < |X - 1/2| \leq t) = 4t^2$
5. $f(x) = 12(x - 1/2)^2 \implies \mathbb{P}(0 < |X - 1/2| \leq t) = 8t^3$



Note that $\alpha = 0$ is no assumption at all, since then the margin condition is just saying the probability is bounded by some constant. Therefore the margin assumption is only imposing extra structure when we take $\alpha > 0$. One important and often practical case is $\alpha = 1$, which is implied by $\mu(X)$ having a bounded density, as shown in the following result.

Proposition 2. *If $\mu(X)$ has a density bounded above by $B < \infty$, then Assumption 1 holds with $C = 2B$ and $\alpha = 1$.*

Proof. Letting Q denote the distribution of $\mu(X)$, we have

$$\begin{aligned}\mathbb{P}(0 < |\mu(X) - 1/2| \leq t) &= \int \mathbb{1}(0 < |\mu - 1/2| \leq t)\, dQ(\mu) \\ &\leq B \int \mathbb{1}(1/2 - t \leq \mu \leq 1/2 + t)\, d\mu = 2Bt.\end{aligned}$$

□

The case $\alpha = \infty$ represents the setting where there is not only little mass near $1/2$, there is zero mass, i.e.,

$$\mathbb{P}(0 < |\mu(X) - 1/2| \leq t_0) = 0$$

for some $0 < t_0 \leq 1/2$. In this case one would expect accurate classification to be quite easy; we will show this more formally shortly.

The margin condition does not change the complexity of the class of regression functions, but it does add non-trivial structure that can result in faster rates of convergence, as we will see. However, there can be some tension between the margin parameter α and for example smoothness s of the regression function μ in the following sense. Intuitively, if the regression function is very smooth, it cannot move very quickly away from the problematic point $1/2$. Thus if the covariate density is lower bounded near this point, then the margin condition cannot hold for large α, since this would require the function to move away from $1/2$ too abruptly.

In fact, Audibert and Tsybakov [2007] shows the following result indicating that, if α and s are large enough, then the regression function $\mu(x)$ cannot cross $1/2$.

Proposition 3. *Assume:*

1. *the margin condition,*
2. *the regression function $\mu(x)$ is s-smooth,*
3. *the covariate distribution has compact support and a density that is bounded above and below away from zero.*

If $\alpha \min(s, 1) > 1$ (regardless of dimension) then the regression function $\mu(x)$ cannot cross 1/2, i.e., the decision set $\{x : \mu(x) > 1/2\}$ has measure zero or one.

In many real-life examples, we typically expect the regression function to cross $1/2$ and the decision set to be non-trivial; this suggests we should perhaps prioritize $\alpha \leq 1$ when $s \geq 1$.

6 Estimation under the margin assumption

Theorem 7. *Assume the margin condition, and let* $\widehat{h}(x) = \mathbb{1}\{\widehat{\mu}(x) > 1/2\}$ *be the plug-in estimator. Then*

$$R(\widehat{h}) - R(h^*) \leq 2C\|\widehat{\mu} - \mu\|_\infty^{1+\alpha}.$$

Proof. Using the fact that $\widehat{h}(x) \neq h^*(x)$ implies $|\mu(x) - 1/2| \leq |\widehat{\mu}(x) - \mu(x)|$, we have

$$\begin{aligned}
R(\widehat{h}) - R(h^*) &= \int |2\mu(x) - 1| \mathbb{1}\{\widehat{h}(x) \neq h^*(x)\} \, d\mathbb{P}(x) \\
&\leq 2 \int |\widehat{\mu}(x) - \mu(x)| \mathbb{1}\{0 < |\mu(x) - 1/2| \leq |\widehat{\mu}(x) - \mu(x)|\} \, d\mathbb{P}(x) \\
&\leq 2\|\widehat{\mu} - \mu\|_\infty \mathbb{P}(0 < |\mu(X) - 1/2| \leq \|\widehat{\mu} - \mu\|_\infty) \\
&\leq 2C\|\widehat{\mu} - \mu\|_\infty^{1+\alpha}
\end{aligned}$$

where we used the margin condition in the last line. □

This result shows how we can get faster rates via the margin condition, even with simple plug-in estimators. For example, if μ is s-smooth then standard estimators (local polynomials, series estimators) with appropriate tuning satisfy

$$\|\widehat{\mu} - \mu\|_\infty = O_{\mathbb{P}}\left((n/\log n)^{-1/(2+d/s)}\right)$$

and the same in expectation after raising to some power, so that the above then implies

$$\mathbb{E}\{R(\widehat{h}) - R(h^*)\} \leq C(n/\log n)^{-(1+\alpha)/(2+d/s)}.$$

Audibert and Tsybakov [2007] show this rate is almost minimax optimal, up to log factors:

Theorem 8. *Let* $\mathcal{P}$ *denote the set of distributions for which the*

1. *the regression function* $\mu(x)$ *is* s*-smooth,*
2. *the margin condition holds for some* $\alpha \leq d/s$*,*
3. *the covariate distribution has compact support and a density that is bounded above and below away from zero.*

Then

$$\inf_{\widehat{h}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\{R(\widehat{h}) - R(h^*)\} \geq Cn^{\frac{-(1+\alpha)}{2+d/s}}.$$

Can the gap between these upper and lower bounds be resolved? Audibert and Tsybakov [2007] show it can, i.e., that the upper bound can be improved to match the lower bound, if we have an appropriate exponential concentration inequality for the regression estimator.

Theorem 9. *Assume*

$$\mathbb{P}(|\widehat{\mu}(x) - \mu(x)| \geq \delta) \leq C_1 \exp(-C_2 r_n \delta^2)$$

and the margin condition. Then the plug-in estimator $\widehat{h}(x) = \mathbb{1}\{\widehat{\mu}(x) > 1/2\}$ *satisfies*

$$\mathbb{E}\{R(\widehat{h}) - R(h^*)\} \lesssim r_n^{-(1+\alpha)/2}.$$

Proof. For some $\delta > 0$ define the sets

$$\begin{aligned} A_0 &= \{x : 0 < |\mu(x) - 1/2| \leq \delta\} \\ A_j &= \{x : 2^{j-1}\delta < |\mu(x) - 1/2| \leq 2^j \delta\}, \quad \text{for } j \geq 1 \end{aligned}$$

and write

$$\begin{aligned} \mathbb{E}\{R(\widehat{h}) - R(h^*)\} &= \mathbb{E}\Big[|2\mu(X) - 1|\mathbb{1}\{\widehat{h}(X) \neq h^*(X)\}\Big] \\ &= \sum_{j=0}^{\infty} \mathbb{E}\Big[|2\mu(X) - 1|\mathbb{1}\{\widehat{h}(X) \neq h^*(X)\}\mathbb{1}(X \in A_j)\Big]. \end{aligned}$$

For $j = 0$ we have by the margin condition that

$$\begin{aligned} \mathbb{E}\Big[|2\mu(X) - 1|\mathbb{1}\{\widehat{h}(X) \neq h^*(X)\}\mathbb{1}(X \in A_0)\Big] &\leq \mathbb{E}\Big[|2\mu(X) - 1|\mathbb{1}(0 < |\mu(X) - 1/2| \leq \delta)\Big] \\ &\leq 2\delta \mathbb{P}(0 < |\mu(X) - 1/2| \leq \delta) \leq 2C\delta^{1+\alpha}. \end{aligned}$$

For $j \geq 1$, again using that $|\mu - 1/2| \leq |\widehat{\mu} - \mu|$ whenever $\widehat{h} \neq h^*$, we have

$$\begin{aligned} &\sum_{j\geq 1} \mathbb{E}\Big[|2\mu(X) - 1|\mathbb{1}\{\widehat{h}(X) \neq h^*(X)\}\mathbb{1}(X \in A_j)\Big] \\ &\quad = \sum_{j\geq 1} \mathbb{E}\Big\{|2\mu - 1|\mathbb{1}(0 < |\mu - 1/2| \leq |\widehat{\mu} - \mu|)\mathbb{1}(2^{j-1}\delta < |\mu - 1/2| \leq 2^j\delta)\Big\} \\ &\quad \leq \sum_{j\geq 1} 2^{j+1}\delta \mathbb{E}\Big\{\mathbb{1}(2^{j-1}\delta \leq |\widehat{\mu} - \mu|)\mathbb{1}(0 < |\mu - 1/2| \leq 2^j\delta)\Big\} \\ &\quad = \sum_{j\geq 1} 2^{j+1}\delta \int \mathbb{P}(|\widehat{\mu}(x) - \mu(x)| \geq 2^{j-1}\delta)\mathbb{1}\{0 < |\mu(x) - 1/2| \leq 2^j\delta\}\, d\mathbb{P}(x) \\ &\quad \leq \sum_{j\geq 1} 2^{j+1}\delta C_1 \exp\{-C_2 r_n (2^{j-1}\delta)^2\} C(2^j\delta)^\alpha \\ &\quad = 2CC_1\delta^{1+\alpha} \sum_{j\geq 1} 2^{j(1+\alpha)} \exp\{-C_2 r_n (2^{j-1}\delta)^2\} \end{aligned}$$

Now take $\delta = r_n^{-1/2}$ to obtain

$$\mathbb{E}\{R(\widehat{h}) - R(h^*)\} \lesssim r_n^{-(1+\alpha)/2} + r_n^{-(1+\alpha)/2} \sum_{j\geq 1} 2^{j(1+\alpha)} \exp(-C_2 2^{2j-2}) \lesssim r_n^{-(1+\alpha)/2}$$

as desired. □

Audibert and Tsybakov [2007] also shows that standard local polynomial estimators satisfy the exponential inequality with

$$\mathbb{P}(|\widehat{\mu}(x) - \mu(x)| \geq \delta) \leq C_1 \exp(-C_2 n h^d \delta^2)$$

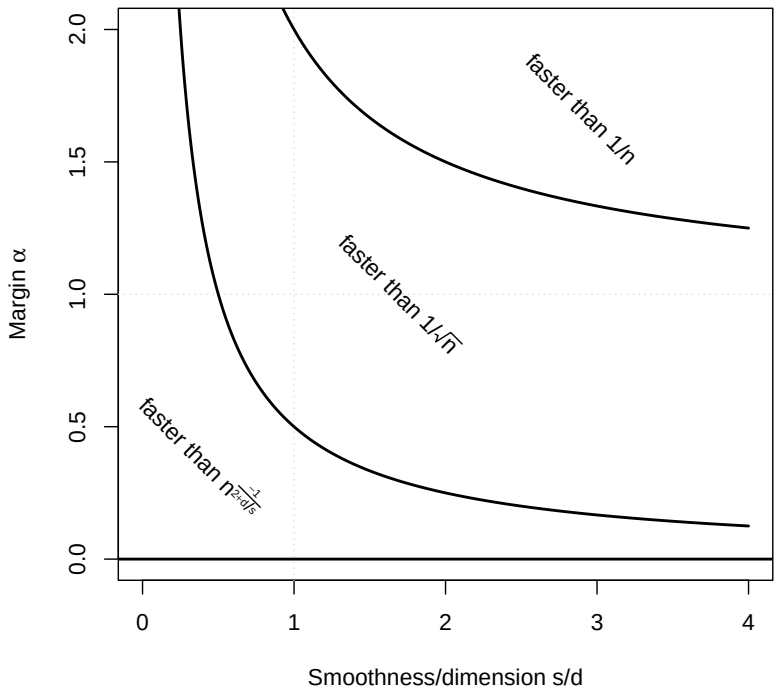
so that if $h \sim n^{-1/(2s+d)}$ then

$$\mathbb{P}(|\widehat{\mu}(x) - \mu(x)| \geq \delta) \leq C_1 \exp(-C_2 n^{2s/(2s+d)} \delta^2)$$

and one obtains the minimax rate

$$n^{-\frac{2s}{2s+d}(1+\alpha)/2} = n^{-\frac{(1+\alpha)}{2+d/s}}.$$

Note that this rate matches that of smooth nonparametric regression in the most difficult case when $\alpha = 0$ (i.e., when $\mu(X)$ could have arbitrary mass at/near the problematic point $1/2$), but otherwise it is faster. In fact when $\alpha > d/2s$ the rate is faster than the parametric rate $1/\sqrt{n}$, and when $\alpha > d/s + 1$ the rate is super-fast, i.e., faster than $1/n$. However recall that in this super-fast regime the decision set has measure zero or one since the regression function cannot cross $1/2$. The rates are illustrated in the plot below.



Audibert and Tsybakov [2007] goes on to show that in the very favorable case where there is zero mass around $1/2$, i.e., that $\alpha = \infty$ so $\mathbb{P}(0 < |\mu(X) - 1/2| \leq t_0) = 0$ for some t_0, plug-in estimators yield *exponentially fast* convergence rates, of the order $1/\exp(n)$. A notable distinction between plug-in estimators and ERM is that the latter cannot attain faster than $1/n$ rates even here, thanks to a lower bound result from Massart and Nedelec [2006].

References

J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

P. Massart and E. Nedelec. Risk bounds for statistical learning. *Annals of Statistics*, 34(5): 2326–2366, 2006.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Y. Yang. Minimax nonparametric classification. i. rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999.