Representations: linear, linear with an expressive basis.

We started class with the following administrative issues:

- I asked the CS staff for help with a room, and even reached out to ECE myself; no room for now =(. Post in the forum if you are getting back pain or something like that from standing/sitting, we'll find a solution.
- The enrollment cap may increase; I'm still trying to figure out how I'd achieve fair and useful grading.
- From now on, we'll prove what we can and push unfinished material to the next class; no more rushing.
- Class on September 28 will probably be canceled; please consider attending Allerton that day!
- I explained "loss" and "risk" again, and pointed out that loss is sometimes written with 2 parameters, for instance $\ell(y, \hat{y}) = (y - \hat{y})^2$. Often, ℓ is not symmetric and order matters. There are many conventions, feel free to ask about this stuff.
- It may be useful to ctrl-f on the lecture notes for "open" and "project".

Recall: the simplest setup in **statistical learning theory** asks us to minimize **risk**, for instance the least squares risk, $\mathcal{R}(f) := \mathbb{E}(f(X) - Y)^2$, over all choices f in some function class $\mathcal{F}$.

Representation question: how much is lost by searching over $\mathcal{F}$ rather than something else?

One concrete formulation:

$$\sup_{g \text{ continuous}} \inf_{f \in \mathcal{F}} \underbrace{\int_{[0,1]^d} |f(x) - g(x)| dx}_{=:\|f-g\|_1} \stackrel{\leq}{\stackrel{?}{\geq}} ??? \ . \tag{1}$$

In words: a lower bound (or upper bound) on how well the first class (continuous functions) is approximated by the second ($\mathcal{F}$).

Remarks.

1. Arguably the right question, given a distribution μ over (X, Y), is to look at

$$\inf_{f \in \mathcal{F}} \int |f(x) - y| d\mu(x, y). \tag{2}$$

   On the other hand, for a fixed target function g, the form in (1) can be converted into this one by considering distributions where $\Pr[Y = g(x)|X = x] = 1$.

2. There are many variations on (1), for instance changing the notion of distance, the notion of the bigger class (e.g., measurable rather than continuous), more generic prediction problems (not just classification and regression), etc. I'll give some references later in today's lecture notes.

3. The shorthand $\|f - g\|_1$ is very convenient. If it looks strange or if you are very curious about it, note it is borrowed from the theory of L^p spaces in functional analysis (e.g., Folland 1999, Chapter 6).

4. **Why is representation important?** I think it's important to revisit this question since these lectures do not appear in other learning theory courses. My personal view is that representation is a signficant part of the success of neural networks.

Open problem. Of course, (1) and (2) are still not quite right: what we really care about is how well we can model *natural functions* (and natural probability distributions). This is a massive question, much of which is arguably non-mathematical.

Linear representations

Linear representations alone can perform poorly.

Theorem (Minsky and Papert (1969, Chapter 0))**.** Consider the *XOR problem*: X is uniform on the set $S := \{(a, b) : a, b \in \{0, 1\}\}$, and define a target function $g(x) := \mathbf{1}[x_1 + x_2 = 1]$. Additionally define the linear classifiers as $\mathcal{L} := \{x \mapsto \mathbf{1}[a^\top x \geq b] : (a, b) \in \mathbb{R}^2 \times \mathbb{R}\}$. Then

$$\inf_{f \in \mathcal{L}} \frac{1}{4} \sum_{x \in S} |f(x) - g(x)| \geq \frac{1}{4}.$$

Remarks.

- This is an expected distance just as before, but we're using a distribution over a discrete set.
- Following convention, these "linear classifiers" are called "linear" despite involving affine functions, and "classifiers" because they output 0 or 1.
- For any boolean predicate P, the notation $\mathbf{1}[P(x)]$ (and $\mathbf{1}_P$) means

$$\mathbf{1}[P(x)] := \mathbf{1}_P(x) := \begin{cases} 1 & \text{when } P(x) \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Fix any linear classifier $f(x) = \mathbf{1}[a^\top x \geq b]$. This classifier corresponds to a halfspace $H := f^{-1}(\{+1\}) = \{x \in \mathbb{R}^2 : a^\top x \geq b\}$ labeled $+1$, and the complement H^c labeled 0. Then *(picture drawn in class)*

$$\begin{aligned} f \text{ has 0 error} &\iff \{(0, 1), (1, 0)\} \subseteq H \wedge \{(0, 0), (1, 1)\} \subseteq H^c \\ &\iff \text{conv}(\{(0, 1), (1, 0)\} \subseteq H \wedge \text{conv}(\{(0, 0), (1, 1)\}) \subseteq H^c, \end{aligned}$$

where $\text{conv}(\cdot)$ denotes the convex hull of a set. But the last statement can not hold since $\text{conv}(\{(0, 1), (1, 0)\})$ and $\text{conv}(\{(0, 0), (1, 1)\})$ are not disjoint whereas H and H^c are disjoint, and therefore "f has 0 error" is also false. Consequently f is wrong on at least 1 of 4 points. Since $f \in \mathcal{L}$ was arbitrary, the bound on the infimum follows. □

Thus linear classifiers alone are inadequate. There is a fix, however: consider linear functions over some *basis* $\mathcal{B}$, meaning functions of the form

$$\text{span}(\mathcal{B}) := \left\{ x \mapsto \sum_{f \in S} a_f f(x) : S \subseteq \mathcal{B}, |S| < \infty, (a_f)_{f \in S} \in \mathbb{R}^{|S|} \right\}.$$

This is the approach used by many machine learning methods, like boosting and SVMs, which are still essentially learning linear functions, but meanwhile able to fit anything.

An example good basis is the set of *monomials*.

Theorem (Weierstrass (1885))**.** If $\mathcal{B}$ consists of all monomial functions (functions of the form $x \mapsto \prod_{i \in S} x_i^{a_i}$ where $S \subseteq \{1, \ldots, d\}$ and a_i is a positive integer), then

$$\sup_{g \text{ continuous}} \inf_{f \in \text{span}(\mathcal{B})} \|f - g\|_1 = 0.$$

There is an incredibly slick proof for this that will appear as a future homework problem (the "magic trick" needed for the proof will be provided).

On the other hand, learning over polynomials is fairly uncommon (certainly over polynomials of arbitrary degree; notice that their cardinality is exponential in dimension, which is bad). Instead, consider the basis of indicators on rectangles: functions $x \mapsto \mathbf{1}[x \in R]$, where R is a rectangle, meaning a product of intervals $R = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d]$. Such functions are common; indeed, they will be used to show approximation properties of linear combinations of decision trees ("boosted decision trees") and neural networks in future lectures.

Lemma. Let continuous $g : \mathbb{R}^d \to \mathbb{R}$ and any $\epsilon > 0$ be given.

1. There exists a partition of $[0,1]^d$ into rectangles $(R_1, \ldots, R_N)$ and a function $h(x) := \sum_{i=1}^N g_i \mathbf{1}[x \in R_i]$ where $g_i := g(x)$ for some $x \in R_i$ such that $\|g - h\|_1 \leq \epsilon/2$. (Without loss of generality, this partition can be the uniform gridding of $[0,1]^d$ into cubes of equal volume.)
2. Let a basis $\mathcal{B}$ given so that for any $\tau > 0$ and rectangle R, there exists $f_R \in \text{span}(\mathcal{B})$ with $\|f_R - \mathbf{1}_R\|_1 \leq \tau$. Then there exists $f \in \text{span}(\mathcal{B})$ with $\|g - f\|_1 \leq \epsilon$.

Proof. This proof will use the following facts from analysis, which are consequences of f being continuous and the relevant domain $[0,1]^d$ being compact. (These are explicit theorems in the excellent introductory analysis book by Rudin (1976, Theorems 4.15 and 4.19); I highly recommend this book, as well as the standard graduate analysis text by Folland (1999).)

- We can choose a tiny $\delta > 0$ so that every $x, x' \in [0,1]^d$ with $\|x - x'\|_\infty < \delta$ satisfy $|g(x) - g(x')| \leq \epsilon/2$.
- Set $M := \sup_{[0,1]^d} |g(x)|$; compactness and continuity imply $M < \infty$.

The proofs of the two parts are as follows. *(In class, a picture was drawn with a function $g : \mathbb{R}^2 \to \mathbb{R}$ being approximated via the forthcoming function h which is constant over each cell in a uniform grid partition of $[0,1]^d$.)*

1. Take $\delta > 0$ as in the first analysis fact, and pick any integer $k \geq 1/\delta$. Let $(R_1, \ldots, R_N)$ with $N := k^d$ denote the uniform partition of $[0,1]^d$ into disjoint (sometimes half-open) cubes, and for each i set $g_i := g(x)$ for an arbitrary $x \in R_i$. Then $h(x) := \sum_{i=1}^n g_i \mathbf{1}[x \in R_i]$ satisfies

$$\int_{[0,1]^d} |h(x) - g(x)| dx = \sum_{i=1}^N \int_{R_i} |h(x) - g(x)| dx = \sum_{i=1}^N \int_{R_i} |g_i - g(x)| dx \leq \sum_{i=1}^N \int_{R_i} \frac{\epsilon}{2} dx = \frac{\epsilon}{2}.$$

2. Let h be as given in the previous section. Next, for every R_i, choose $f_i \in \text{span}(\mathcal{B})$ such that $\|f_i - \mathbf{1}_{R_i}\|_1 \leq \tau$, where τ will be specified at the end of the proof. Then the function $f := \sum_i g_i f_i \in \text{span}(\mathcal{B})$ satisfies

$$\|f - g\|_1 \leq \|f - h\| + \|h - g\|_1 \leq \int_{[0,1]^d} \left| \sum_i g_i f_i(x) - \sum_i g_i \mathbf{1}[x \in R_i] \right| dx + \epsilon/2$$

$$\leq \sum_i \int_{[0,1]^d} |g_i| \left| f_i(x) - \mathbf{1}[x \in R_i] \right| dx + \epsilon/2 \leq NM\tau + \epsilon/2.$$

The proof is complete by choosing $\tau := \epsilon/(2NM)$. $\square$

Remarks.

1. These proofs exhibit a *curse of dimension*: the number of basis elements used is exponential in dimension. Very bad.
2. There is a tiny bit of cheating in the proof: if the cubes are to form a partition, they must be disjoint, and thus some will be the products of half-open intervals $[i/k, (i+1)/k)$ where $i+1 < k$. The resolution is to allow "rectangles" and "cubes" to be produces of closed, half-open, and open intervals, which does not change any volume computations and indeed grants the correctness of all above expressions.
3. Regarding the aforementioned issue of representation questions over more general measures; unfortunately I only know of one result here. No doubt others exist, and I am happy to hear of them. It is unfortunately a reference to one of my own papers; on the other hand, I include a fair bit of discussion of similar results. Anyway (Telgarsky 2013, Appendix B).

References

Folland, Gerald B. 1999. *Real Analysis: Modern Techniques and Their Applications*. 2nd ed. Wiley Interscience.

Minsky, Marvin, and Seymour Papert. 1969. *Perceptrons: An Introduction to Computational Geometry*. MIT Press.

Rudin, Walter. 1976. *Principles of Mathematical Analysis*. 3rd ed. McGraw-Hill Book Company.

Telgarsky, Matus. 2013. "Boosting with the Logistic Loss Is Consistent." In *COLT*.

Weierstrass, Karl. 1885. "Über Die Analytische Darstellbarkeit Sogenannter Willkürlicher Functionen Einer Reellen Veränderlichen." *Sitzungsberichte Der Akademie Zu Berlin*, 633–39, 789–805.