Basics of statistical learning terminology.

1 Definitions

The fundamental problem of "statistical learning" can be described as follows: we possess a data set made of n "features" $X_1, \dots, X_n$ and n associated "labels". The goal is to develop a rule in order, given a new feature X_{n+1}, to predict the "correct" label Y_{n+1}.

We will always work under the assumption that the couples (Feature, Label) given by $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent and that they have a common distribution P. Note that the independence assumption is a little naive in some contexts. We note Features $\subset \mathbb{R}^m$ the space of features and Labels $\subset \mathbb{R}^n$ the space of labels.

There are plenty of examples:

1. X is the couple (Age, Gender) and Y is the Size.
2. X is the three SAT scores and Y is the GPA.
3. X are the pixels of a picture and Y is the Yes/No answer to "Is it a cat?"
4. X is a collection of informations on a city: the weather, the time of the year, whether there is an important event or not, the amount of people who are looking at it on your travel agency website, etc. and Y is the price at which you can sell a flight ticket.

Situations 1. and 2. are "low-dimensional", in comparison to 4. and especially to 3. where the feature space has very high dimension.

Also, in situation 2. we expect the relationship between Y and X to be very simple. In 1. and to certain extent in 4. we can imagine a not-too-complicated dependency. In 3, however, the relationship seems very complicated.

Predictors A **predictor** is a function from Features to Labels. How can we measure the quality of a predictor? We introduce a cost function

$$c : \text{Labels} \times \text{Labels} \to \mathbb{R}$$

that measures the distance between two labels. This is something **we choose**. Natural examples include

- The L^2-cost or quadratic cost: if $y_1 = (y_{1,1}, \dots, y_{1,n})$ and $y_2 = (y_{2,1}, \dots, y_{2,n})$ are two labels,

$$L^2\text{-cost} \ = \left(\sum_{k=1}^{n} (y_{1,k} - y_{2,k})^2 \right)^{1/2}$$

- The L^1-cost

$$L^1\text{-cost } = \sum_{k=1}^{n} |y_{1,k} - y_{2,k}|$$

For a fixed cost c, we define the risk of a predictor f as

$$\mathcal{R}(f) := \mathbb{E}_P \left[c\left(f(X), Y\right) \right].$$

It gives the average (under P) of the "distance" (as measured by the cost) between the predicted label $f(X)$ and the "real" label Y. We may now rephrase our goal as follows:

Goal: given a cost, and a data set, find a predictor that minimizes the risk.

Of course, since P is in general unknown, we will need to consider a slightly different question. But for a moment, let us assume we are in an ideal setting where we know P, not just a data set. We define

- The Bayes risk $\mathcal{R}^* = \inf_{f \text{predictor}} \mathcal{R}(f)$, the minimal risk over all predictors.
- A Bayes predictor as a predictor f^* such that $\mathcal{R}(f^*) = \mathcal{R}^*$, that is a predictor with minimal risk.

Learning rules In practice, we will build our prediction based on the data. So we design a "learning rule" : $\hat{f}$: Data $\times$ Features $\to$ Labels. For a given data set D_n

$$\hat{f}(D_n, \cdot) : \text{Features} \to \text{Labels}$$

is a predictor, in the previous sense.

There are two ways to measure the risk of this learning rule:

- If we first sample the data set D_n and consider

$$\mathcal{R}\left(\hat{f}(D_n, \cdot)\right) := \mathbb{E}\left[c(\hat{f}(D_n, X), Y) | D_n\right]$$

  which is the *risk* of the predictor $\hat{f}(D_n, \cdot)$. It is a random variable when we consider D_n as random.

- If we take the average of the previous quantity over all possible data sets, we obtain the "average risk"

$$\mathbb{E}\left[\mathcal{R}\left(\hat{f}(D_n, \cdot)\right)\right] = \mathbb{E}\left[c(\hat{f}(D_n, X), Y)\right]$$

A "good" learning rule is such that, for example, as $n \to \infty$

- The risk converges to the Bayes risk (almost surely).

- The average risk converges to the Bayes risk.

Important remark: in practice, we cannot compute the risk of a learning rule because... we don't know P! As usual, we resort to an *empirical* mindset and compute an "empirical risk" based on data. Which leads to an obvious danger: if we compute the risk based on the *same* data as the one we used to build the learning rule, we will do extremely well, but in a meaningless way. These are the (very) important topics of *overfitting*, *generalization* etc. that we will not consider in this class. Just remember to be careful.

Regression and classification Usually, we say that we are in a "regression" situation when the values of the labels are continuous (e.g. a size, a price, a temperature) and "classification" when the labels belong to a discrete space (e.g. Yes/No, 0/1, North/West/East/South).

Beware that the terminology can vary. Also, it is sometimes useful to treat certain classification problems as regression problems, e.g. by seeing a 0/1 label as a label on $[0, 1]$.

2 Some examples

Deterministic Let us start with situations where the label is a deterministic function $Y = r(X)$ of the feature. In this case, the Bayes risk is always zero. However, we may encounter very different cases:

- Completely disordered: if r is a function with no regularity, then the task of learning r from the data is impossible.
- Dictatorial: if r is a constant, then it is enough to observe **one** data point in order to find the perfect learning rule.
- Anything in between. We may for example think of a linear dependency $Y = aX + b$ (here $m = n = 1$). Then observing **two** data points in enough. Question: what is the analogue result in higher dimension?

Note also that there are plenty of different ways to be "regular". We could assume something like "the opinion of my neighbors influence my opinion", which could be turned into a mathematical statement of the type "$r(x)$ is close to $\frac{1}{2}(r(x-1) + r(x+1))$... These are *modelling* considerations, and different models yield different learning rules.

Noise We now study a "signal plus noise" situation, where the distribution of the label, knowing the feature X, is of the form

$$Y = r(X) + \varepsilon\mathcal{N}(0, 1)$$

We call $r(X)$ the "signal" and $\varepsilon\mathcal{N}(0, 1)$ the "noise".

What becomes of the Bayes risk? In this case, even if we know the distribution we cannot avoid the noise. Depending on the cost function, it might be optimal (i.e. the Bayes predictor wants) to answer $r(X)$ as a prediction for the label of X, in which case the Bayes risk will be

$$R^* = \mathbb{E}\left[c(r(X), r(X) + \varepsilon Z)\right], \text{ where } Z \sim \mathcal{N}(0,1)$$

Let us start with a "dictatorial plus noise" case.

$$Y = d + \varepsilon \mathcal{N}(0,1),$$

where d is the "dictatorial constant". What is a good learning rule? Make an observation $(X_1, Y_1), \ldots, (X_n, Y_n)$. We have

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = d + \frac{\varepsilon}{\sqrt{N}}\mathcal{N}(0,1)$$

(Question: why?) We want predict d all the time, which seems like the optimal thing to do (Bayes predictor), and we are thus tempted to predict $\hat{d}_n := \frac{1}{n}\sum_{i=1}^{n} Y_i$. Let us ask the question: when can we guarantee that $\hat{d}_n$ will be between $0.9d$ and $1.1d$, 99% of the time?

$$\mathbb{P}\left[\left|\frac{\varepsilon}{\sqrt{N}}\mathcal{N}(0,1)\right| \leq \frac{d}{10}\right] \geq 0.99 \text{ ??}$$

yields

$$\mathbb{P}\left[|\mathcal{N}(0,1)| \leq \frac{d}{\varepsilon}\frac{\sqrt{n}}{10}\right] \geq 0.99 \text{ ??}$$

which we can read in tables. The quotient $\frac{d}{\varepsilon}$ is (related to) the so-called *signal-to-noise* ratio.

Linear + noise?