Chapter 4

Risk

In order to make sense of the idea of learning from data, we need first an idea of what we are learning from. Everything we do when learning from data is based on assumptions about the data generating process. When we write down all the assumptions about the data generating process the *statistical model* is all the distributions that satisfy those assumptions.

Definition 4.1. *A* **statistical model** *is an indexed family of distributions* $\mathcal{S} = \{f(x;\theta), \theta \in \Theta\}$.

- A **parametric model** is a model where the indexing parameter θ is a vector in k-dimensional Euclidean space. That is, θ is finite dimensional.
- A **non-parametric model** is a model where Θ is infinite dimensional.

Example 4.2. $\mathcal{N} = \left\{ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{|x-\mu|}{\sigma}\right)^2}, \mu \in \mathbb{R}, \sigma > 0 \right\}$. *This is the set of all normal distributions on* $\mathbb{R}$. *It is also a parametric model since there are two parameters* μ *and* σ.

Example 4.3. $\mathcal{E} = \{F : F \text{ is a CDF}\}$. *This is a non-parametric model, as the set of all CDFs is infinite dimensional. What this means is that there is no finite set of parameters that can describe all CDFs.*

To reiterate, a statistical model is a model of the data generation, that is, it is what we assume the truth is. As you can see above, a parametric model is more restrictive, this usually means that drawing conclusions (estimation) from data is "easier" (higher precision with less data).

Let us quote Merriam-Webster:

"Main Entry: **in·fer·ence**
Pronunciation: 'in-f(&-)r&n(t)s, -f&rn(t)s
Function: *noun*

Date: 1594

1. the act or process of inferring: as

   (a) the act of passing from one proposition, statement, or judgment considered as true to another whose truth is believed to follow from that of the former

   (b) the act of passing from statistical sample data to generalizations (as of the value of population parameters) usually with calculated degrees of certainty."

Inference lies at the heart of statistics and learning. The question is: what do we want to know?

Under a statistical model $\mathcal{F}$, there is a hidden $f^* \in \mathcal{F}$ that generates the data, we would like to infer something about f^* using observations.

Here are some examples of inference problems:

1. Density estimation, or consequences of the density, like estimating the probability of an event.

2. Estimating the distribution function. Can be used to answer questions about probabilities of simpler events, but can also be functionals of the distribution.

3. Functional dependence, usually regression, or pattern recognition.

4.1 The supervised learning problem

As we will be working with machine learning and data science let us describe the learning problem as seen from the field of computer science and let us interpret each piece using our probabilistic terminology. We will begin with learning a functional dependency, the model contains three elements

1. The generator of the data G

2. The supervisor S

3. The learning machine LM.

The generator G is a source of situations, we will make the simplest assumptions, that G generates vectors X_i i.i.d. according to some unknown but fixed distribution $F(x)$. These vectors X_i are inputs to the *supervisor* that outputs a value Y_i, we know the supervisor has an unknown function transforming X_i into Y_i. At this point we could also consider that $Y_i|X_i$ has some noise in it (perhaps the supervisor is measuring something about

X_i but that measurement has a random error in it). The learning machine observes a realization of the n pairs

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

we denote this realized set as $(x_1, y_1), \ldots, (x_n, y_n)$ (the training set). In this course we make the assumption that the supervisor generates Y_i from X_i according to an unknown conditional distribution $F(y|x)$, that is the conditional distribution of $Y_i|X_i$. Recall that this includes the case of a functional dependency $y = f(x)$. The learning machine thus observes pairs $(X_i, Y_i) \sim F_{X,Y}(x, y)$ where $f_{X,Y}(x, y) = f_{Y|X}(y|x)f_X(x)$ is the joint distribution.

> The goal is to approximate this functional dependency in some way using the observations!

Whenever we want to produce an approximation it often makes sense to come up with a measure of quality. Denote $z = (x, y)$ and consider a function $g(z)$ that is of a type we are interested in and define a loss functional $L(z, g)$ that measures the quality of g at the point z. Now consider the expected loss, usually denoted as *Risk*:

$$R(g) = \int L(z, g)dF(z) = \mathbb{E}\left[L(Z, g)\right]$$

where $Z = (X, Y) \sim F(x, y)$.

> Goal of the learning machine: Define a class of functions g to search from and minimize the risk inside this class.

We will now work through how this is formulated mathematically in some special cases that will be general enough for us in this course. The purpose, for now, is to get a feeling for the concepts. Later we will move on to, how to actually minimize the risk using empirical data (Empirical Risk Minimization) and some guarantees we can make under certain assumptions.

4.1.1 Mathematical description of the learning problem "find f"

Let us begin by describing the learning problem for a simple case where the supervisor is using a real valued continuous function $y = f(x)$, for $x \in [0, 1]$, i.e. $f \in C([0, 1], \mathbb{R})$, that is $F(y|x) = \mathbb{1}_{y \geq f(x)}(y)$, i.e. a point mass centered at $f(x)$.

Example 4.4. *Let us motivate the above setup with a typical example. In many image analysis problems you need a way to determine the scale of*

the objects in view. This is usually done using a fiducial mark, that is a reference shape of known size. This is often a solid circle of known size.

What we need in order to determine the scale of the image is to detect the fiducial mark and measure how many pixels it corresponds to in the image. Since it is a circle, we only need to figure out the radius of it in the image (in terms of the number of pixels).

- *The data generator G is the process that produces the images, i.e. the experiment. The data that is generated is X which is the image.*
- *The supervisor is anyone or anything that knows the answer of the radius of the circle. That is, if you give the supervisor the image X, the output will be the correct radius of the circle Y. Our assumption is that $Y = f(X)$.*
- *The goal of the learning machine is to figure out how to get from the image X to the radius Y simply by observing examples of pairs (X, Y).*

To describe the learning problem in the above setting we need to set up the following things:

1. Statistical model: $\mathcal{S} = \{F(x, y) = \mathbb{1}_{y \geq f(x)}(y)F(x), f \in C([0, 1], \mathbb{R})\}$
2. Model space: $\mathcal{M} = \{g_\lambda(x) : \lambda \in \Lambda\}$, a parametrized space of functions in which we are searching for f, or an approximation thereof. (The kind of functions the learning machine can represent).
3. A loss function $L : [0, 1] \times \mathbb{R} \to \mathbb{R}$.

In this setting the learning problem becomes the risk minimization problem below

$$g^* = \operatorname*{argmin}_{g \in \mathcal{M}} R(g)$$

However, since the model space is parametrized we can actually rewrite the minimization problem to be over Λ instead, as follows

$$\lambda^* = \operatorname*{argmin}_{\lambda \in \Lambda} R(g_\lambda).$$

This also allows us to write the loss function as a function of z and the parameter λ as $L(z, g_\lambda(z)) = Q(z, \lambda)$ and our risk can be written as

$$R(\lambda) = \int Q(z, \lambda) dF(z) = \int Q((x, f(x)), \lambda) dF(x)$$

where the first integral is a double integral over $z = (x, y)$ and the second integral is a single integral over x.

As you have seen above the loss function is not specified and can be chosen quite freely. Let us next consider the problem of regression.

4.1.2 Finding the regression function $r(x) = \mathbb{E}[Y|X]$

Let us now assume that the supervisor is generating Y from $F(y|x)$ given the value of X. In this setup, perhaps we would like to estimate the full conditional distribution $F(y|x)$, but this is a hard problem. Instead one could try to estimate some of its properties, for instance, we could try to estimate a functional of $F(y|x)$. The concept of regression is that we are interested in the following functional

$$r(x) = \int y dF(y|x) = \mathbb{E}[Y \mid X = x].$$

Here, r is called the regression function. Let us assume that $Y \in L^2(\mathbb{P})$ and $r \in L^2(dF_X)$, that is

$$\mathbb{E}[Y^2] < \infty, \quad \mathbb{E}[r^2(X)] < \infty.$$

Example 4.5. *Finding the regression function is an extension of the "finding the function", therefore we could use the same example with the fiducial point, but now we could assume that the supervisor doesn't know the exact size of the circle. But instead is performing a measurement which has some error connected to it.*

Example 4.6. *We have already seen an example of a regression problem. Namely Section 2.7.2. Here the goal was to estimate*

$$F_{X|Z}(x \mid z)$$

if we change this to the notation above, the Z is the data coming from our data-generator and the supervisor gives us X sampled from $F_{X|Z}(x \mid z)$.

In this case the statistical model is

$$\mathcal{S} = \Big\{ F(x,y) = F_{Y|X}(y \mid x)F(x); \\ r(x) = \int y dF(y \mid x), \mathbb{E}[r(X)^2] < \infty, \mathbb{E}[Y^2] < \infty \Big\}$$

Consider now an model space $\mathcal{M} = \{g_\lambda(x) : \lambda \in \Lambda, g_\lambda \in L^2(dF_X)\}$ of some functions g_λ parametrized by λ.

For a function $g_\lambda \in \mathcal{M}$ we consider the following risk

$$R(\lambda) = \int (y - g_\lambda(x))^2 dF(x,y)$$

WARNING: we have not specified if $r \in \mathcal{M}$!!

Assume there exists a $\lambda^* \in \Lambda$ such that for $g^* = g_{\lambda^*}$

$$R(g^*) = \inf_{g \in \mathcal{M}} R(g)$$

then write the risk as

$$\begin{aligned} R(\lambda) &= \mathbb{E}\left[(Y - g_\lambda(X))^2\right] = \mathbb{E}\left[(Y - g_\lambda(X) + r(X) - r(X))^2\right] \\ &= \mathbb{E}\left[(Y - r)^2\right] + \mathbb{E}\left[(r(X) - g_\lambda(X))^2\right] + 2\,\mathbb{E}\left[(Y - r)(r - g_\lambda)\right] \\ &= I + II + III. \end{aligned}$$

The assumption that Y, $r(X)$ and $g_\lambda(X)$ all have finite second moment is what allows us to do the computation above, and know that all terms involved are finite. Let us now consider III and note that by the tower property and the definition of r that

$$\begin{aligned} III &= 2\,\mathbb{E}\left[(Y - r)(r - g_\lambda)\right] = 2\,\mathbb{E}\left[\,\mathbb{E}\left[(Y - r)(r - g_\lambda)|X\right]\right] \\ &= 2\,\mathbb{E}\left[(\mathbb{E}\left[Y|X\right] - r)(r - g_\lambda)\right] = 0 \end{aligned}$$

From this we now see that

$$\operatorname*{argmin}_{\lambda \in \Lambda} R(\lambda) = \operatorname*{argmin}_{\lambda \in \Lambda} \mathbb{E}\left[(r(X) - g_\lambda(X))^2\right]$$

which means that the minimizer g^* will be the function in $\mathcal{M}$ that is closest to r in the mean square sense (or in L^2 if using function space notation).

NOTE: if $r \in \mathcal{M}$ then $g^* = r$ a.e. with respect to dF_X.

4.1.3 The pattern recognition problem (classification)

In the pattern recognition model we assume that the supervisors conditional distribution $F(y|x)$ is discrete, and can take k different values, $y = 0, \ldots, k-1$. Consider a model space $\mathcal{M} = \{g_\lambda(x) : g_\lambda(x) \in \{0, \ldots, k-1\}\}$, that is, functions g_λ that takes values in $\{0, \ldots, k-1\}$. It is common to call the functions in the pattern recognition problem, g_λ a **decision function** or **decision rule**. With this at hand, we define the $0-1$ loss function for $z = (x, y)$

$$L(z, u) = \begin{cases} 0 & \text{if } y = u \\ 1 & \text{if } y \neq u \end{cases}$$

that is, the loss is 1 if u is the wrong value and 0 if it is correct. The pattern recognition problem is the problem of minimizing the functional

$$R(\lambda) = \int L(y, g_\lambda(x)) dF(x, y) = \mathbb{E}\left[L(Y, g_\lambda(X))\right]$$

where $(X, Y) \sim F(x, y)$.

Exercise 4.7. *What is a reasonable statistical model for the Pattern Recognition problem?*

The classification problem is in modern times very often associated with the prototypical example of classification of images of dogs and cats. In that example, the image is X and the class is given by Y.

The risk above has a natural interpretation, given the "decision rule" g_λ, the risk $R(\lambda)$ is the probability of an incorrect classification by the rule g_λ,

$$\mathbb{E}\left[L(Y, g_\lambda(X))\right] = \mathbb{P}(\{Y \neq g_\lambda(X)\}).$$

Bayes rule

What is the optimal decision rule? Recall that in the regression setting we had the regression function as the minimizer, but what is it in the pattern recognition problem? Consider the case when $k = 2$ and denote

$$r(x) = \mathbb{E}\left[Y \mid X = x\right] = \mathbb{P}(Y = 1 \mid X = x)$$

Definition 4.8. *The* **Bayes classification rule** h^* *is*

$$h^*(x) = \begin{cases} 1 & \text{if } r(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Let us prove that the Bayes classification rule is the rule that optimizes the risk in the pattern recognition problem.

Theorem 4.9. *For any decision function $g(x)$ taking values in $\{0, 1\}$, we have*

$$R(h^*) \leq R(g).$$

Proof. Note that we can write

$$R(g) = \mathbb{E}\left[L(Y, g(X))\right] = \mathbb{E}\left[\,\mathbb{E}\left[L(Y, g(X)) \mid X\right]\right]$$

we will work only with the inner part, i.e. now

$$\begin{aligned}
\mathbb{E}\left[L(Y, g(X)) \mid X = x\right] &= 1 - \mathbb{E}\left[\mathbb{1}_{\{y=g(x)\}} \mid X = x\right] \\
&= 1 - \mathbb{E}\left[\mathbb{1}_{\{1=g(x)\}}\mathbb{1}_{\{y=1\}} + \mathbb{1}_{\{0=g(x)\}}\mathbb{1}_{\{y=0\}} \mid X = x\right] \\
&= 1 - \mathbb{1}_{\{1=g(x)\}}\,\mathbb{E}\left[\mathbb{1}_{\{y=1\}} \mid X = x\right] - \mathbb{1}_{\{0=g(x)\}}\,\mathbb{E}\left[\mathbb{1}_{\{y=0\}} \mid X = x\right] \\
&= 1 - \mathbb{1}_{\{1=g(x)\}} r(x) - \mathbb{1}_{\{0=g(x)\}}(1 - r(x))
\end{aligned}$$

Now

$$\begin{aligned}
&\mathbb{E}[L(Y,g(X)) \mid X=x] - \mathbb{E}\left[L(Y,h^*(X)) \mid X=x\right] = \\
&\quad = -\mathbb{1}_{\{1=g(x)\}} r(x) - \mathbb{1}_{\{0=g(x)\}}(1-r(x)) + \mathbb{1}_{\{1=h^*(x)\}} r(x) + \mathbb{1}_{\{0=h^*(x)\}}(1-r(x)) \\
&\quad = r(x)(\mathbb{1}_{\{1=h^*(x)\}} - \mathbb{1}_{\{1=g(x)\}}) + (1-r(x))(\mathbb{1}_{\{0=h^*(x)\}} - \mathbb{1}_{\{0=g(x)\}}) \\
&\quad = r(x)(\mathbb{1}_{\{1=h^*(x)\}} - \mathbb{1}_{\{1=g(x)\}}) - (1-r(x))(\mathbb{1}_{\{1=h^*(x)\}} - \mathbb{1}_{\{1=g(x)\}}) \\
&\quad = (2r(x)-1)(\mathbb{1}_{\{1=h^*(x)\}} - \mathbb{1}_{\{1=g(x)\}}) \geq 0.
\end{aligned}$$

This immediately implies the statement of the theorem. □

From the above proof we see that if we are minimizing the risk inside a class $\mathcal{M}$ that does not include h^*, we can write g^* as a minimizer in $\mathcal{M}$ and get

$$R(g^*) = R(h^*) + \mathbb{E}\left[(2r(X)-1)(\mathbb{1}_{\{1=h^*(X)\}} - \mathbb{1}_{\{1=g^*(X)\}})\right]$$

Remark 4.10. *The above expression is interesting.*

- *If $h^*(x) = 1$, which means $r(x) > 1/2$, then the cost of misclassifying is $(2r-1)$, which means that the cost is higher for higher values of r.*
- *If $h^*(x) = 0$, which means that $r(x) \leq 1/2$, then the cost of misclassifying is higher for r close to 0.*
- *In the case $r = 1/2$ there is always zero cost and it does not matter if we misclassify.*

4.2 Maximum Likelihood Estimation

We will now derive the Maximum Likelihood as a special case of risk minimization.

Assume that we have a parametric model $\mathcal{S} = \{p_\alpha(z), \alpha \in \mathbb{R}^d\}$, for some given family of densities p_α. For example we can take

$$p_\alpha(z) = \frac{1}{\sqrt{2\pi\alpha_2}} e^{-\frac{|z-\alpha_1|^2}{2\alpha_2}},$$

where $\alpha = (\alpha_1, \alpha_2)$ which is the Gaussian family. Assume that our underlying model is given by a hidden parameter α^*, then consider the loss function $L(z,\alpha) = -\ln p_\alpha(z)$ then the risk becomes

$$R(\alpha) = -\int \ln(p_\alpha(z)) p_{\alpha^*}(z) dx$$

If we let Z be a random variable with law p_{α^*} then we can write the above as

$$R(\alpha) = \mathbb{E}\left[-\ln(p_\alpha(Z))\right]$$

Given a sequence of i.i.d. random variables $Z_1, \ldots, Z_n$ sampled from p_{α^*} the empirical Risk just becomes

$$\hat{R}(\alpha) = -\frac{1}{n}\sum_{i=1}^{n} \ln(p_\alpha(Z_i)).$$

This is nothing but the negative log likelihood of the observations $Z_1, \ldots, Z_n$ under the model p_α. Thus to minimize the risk with respect to α is the same as maximizing the log likelihood.

So, is the choice of loss L any good? Can we say that the minimum is attained at α^*?

How do we prove that? Well, we prove it using Jensen's inequality (Lemma 2.51). Consider the function $\psi(u) = \ln(u)$ (concave, Jensen is reversed) and $\Phi(x) = \frac{p_\alpha(x)}{p_{\alpha^*}(x)}$, using Jensen's inequality we get

$$R(\alpha^*) - R(\alpha) = \int \psi(\Phi(x)) p_{\alpha^*}(x)dx \leq \psi\left(\int \Phi(x) p_{\alpha^*}(x)dx\right) = \ln 1 = 0$$

as such we have that $R(\alpha^*) \leq R(\alpha)$ and hence α^* is the global minimum of the risk. Is there any other α that also minimizes the risk? We have

$$\int \psi(\Phi(x)) p_{\alpha^*}(x)dx = 0$$

this implies that $\psi(\Phi(x)) = 0$ a.e. with respect to p_{α^*}, so if our family is well behaved (identifiable) then the minimum is unique.

4.2.1 Maximum Likelihood and regression

Suppose that we have a pair (X, Y) of random variable. Denote a proposal joint density of (X, Y) as $f_{X,Y}$. Consider a sequence of i.i.d. samples (X_i, Y_i), $i = 1, \ldots, n$ with the same law as (X, Y), then the negative log-likelihood (which is just the empirical risk under loss ln, see Section 4.2) is given by

$$-\sum_{i=1}^{n} \ln(f_{X,Y}(X_i, Y_i))$$

if we condition on X we get

$$\begin{aligned} -\sum_{i=1}^{n} \ln(f_{X,Y}(X_i, Y_i)) &= -\sum_{i=1}^{n} \ln(f_{Y|X}(Y_i \mid X_i) f_X(X_i)) \\ &= -\sum_{i=1}^{n} \ln(f_{Y|X}(Y_i \mid X_i)) - \sum_{i=1}^{n} \ln(f_X(X_i)) \end{aligned}$$

Now, consider a parametrized family of proposal joint distributions where the marginal density f_X does not depend on any parameter, only the conditional distribution $f_{Y|X}$. Then if we want to minimize the negative log-likelihood over this particular proposal family, only the first summand can change, as such, it is enough to minimize over this. If we flip the sign we get that we would like to maximize the conditional likelihood. This is the main idea behind linear regression and logistic regression, both of which are ubiquitous in the field of data science. To see how this looks like in the context of linear regression, see [W, Chapter 13].

Example 1: Linear regression

In this case we make the assumption that $f_{a,b,\sigma} := f_{a,b,\sigma;Y|X}$ is the density of $N(aX+b, \sigma^2)$ where the parameters of interest are a, b, σ. We assume that f_X is some fixed proposal density, actually it will not matter (see above).

$$-\sum_{i=1}^{n} \ln(f_{Y|X}(Y_i \mid X_i)) = -\sum_{i=1}^{n} \ln\left(\frac{1}{\sigma} e^{-\frac{1}{2\sigma^2}(Y_i-(aX_i+b))^2}\right) - Cn$$

The first term on the right can be rewritten as

$$\sum_{i=1}^{n} \ln(\sigma) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - (aX_i+b))^2$$

The main realization is that minimizing the likelihood gives the same parameters as minimizing the conditional likelihood which gives the same parameters as minimizing the sum of squares. I.e. linear regression in this case is equivalent to mean square regression, as in Section 4.1.2.

Example 2: Logistic regression

Here we assume, on the contrary to linear regression, that the proposal density $f_{\beta_0,\beta_1;Y|X}$ is the density of a Bernoulli$(G(\beta_0+\beta_1 X))$ where the function G is defined as

$$G(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x},$$

and is called the logistic function. Here we assume that $Y \in \{0, 1\}$

If we call $p(X) = G(\beta_0 + \beta_1 X)$ then

$$\begin{aligned} -\sum_{i=1}^{n} \ln(f_{Y|X}(Y_i \mid X_i)) &= -\sum_{i=1}^{n} \ln(p(X_i)^{Y_i}(1-p(X_i))^{1-Y_i}) \\ &= -\sum_{i=1}^{n}(Y_i \ln(p(X_i)) + (1-Y_i)\ln(1-p(X_i))) \end{aligned}$$

Now

$$\ln(p(X_i)) = \ln(1/(1+e^{-(\beta_0+\beta_1 X_i)})) = -\ln(1+e^{-(\beta_0+\beta_1 X_i)})$$
$$\ln(1-p(X_i)) = \ln(1-1/(1+e^{-(\beta_0+\beta_1 X_i)})) = -\ln(1+e^{\beta_0+\beta_1 X_i}).$$

When $Y_i = 0$ we get

$$-\ln(p(X_i)^{Y_i}(1-p(X_i))^{1-Y_i}) = -\ln(1-p(X_i)) = \ln(1+e^{\beta_0+\beta_1 X_i})$$

and when $Y_i = 1$ we get

$$-\ln(p(X_i)^{Y_i}(1-p(X_i))^{1-Y_i}) = -\ln(p(X_i)) = \ln(1+e^{-(\beta_0+\beta_1 X_i)}).$$

Thus the only thing that changes is the sign of the exponent, so if we write $Z_i = 2Y_i - 1$ then $Z_i = 1$ if $Y_i = 1$ and $Z_i = -1$ if $Y_i = 0$ and we can write

$$-\sum_{i=1}^{n} \ln(p(X_i)^{Y_i}(1-p(X_i))^{1-Y_i}) = \sum_{i=1}^{n} \ln(1+e^{-Z_i(\beta_0+\beta_1 X_i)}).$$

Now, you might wonder, why the specific form of $G(x)$ other than the fact that it outputs numbers between 0 and 1? To see why this formula is used, consider the log-odds ratio given X, i.e.

$$\ln\left(\frac{\mathbb{P}(Y=1 \mid X)}{\mathbb{P}(Y=0 \mid X)}\right) = \ln\left(\frac{p(X)}{1-p(X)}\right) = \ln\left(\frac{G(\beta_0+\beta_1 X)}{1-G(\beta_0+\beta_1 X)}\right)$$
$$= \ln(e^{\beta_0+\beta_1 X}) = \beta_0 + \beta_1 X$$

Thus for the logistic regression the log odds ratio is linear.