553.740 Course Notes

August 29, 2023

Contents

1 Lecture 1

1.1 Administrativa

1.1.1 Git Repo

Assignments, starter code, and data will be housed in a Git repository:
https://github.com/schmidttgenstein/fa23-mli/ Please do not push anything to this repo!

1.2 Course Overview

1.2.1 Introduction

From the syllabus:

Machine Learning describes a mishmash of computational techniques for "finding patterns in data." The scope of use, analytic tools, algorithms, and results are almost too numerous to meaningfully batch all such applications under a common appellation. Still, we try. This course focuses on *supervised* machine learning (sml) which roughly deals with using historical *labeled* data to construct a predictor which will correctly label future data.

Formally, we will be operating in space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ denotes "input" space, and $\mathcal{Y}$ "output." While these notions are heuristic, they well frame the situation that we may easily sample data at will from $\mathcal{X}$, while sampling from $\mathcal{Y}$ may be difficult or expensive, and often we would like to decision according to how we believe $x \in \mathcal{X}$ is associated with label $y \in \mathcal{Y}$.

In this course, you will learn how to formulate the supervised learning problem in mathematical terms, how to describe a measure of performance, restrict search space for constructing models for prediction, optimize performance measure in search space, and how to check for generalization. You will learn, also, how to implement some of these methods in code, from the ground up, as well as incorporating pre-built libraries (such as pytorch) for such tasks. Finally, you will learn how to articulate learning guarantees, and understand some of the limits of learning claims. While this course is primarily theory-centric, there will be no dearth of opportunity for employing concrete computational techniques.

1.2.2 The Standard Diagram

Describing our problem space in more detail, consider the following diagram

$$\begin{array}{ccc} \mathcal{X} \times \mathcal{Y} & \xrightarrow{\pi_{\mathcal{Y}}} & \mathcal{Y} \\ \downarrow{\scriptstyle \pi_{\mathcal{X}}} & \nearrow_{\tilde{y}} & \\ \mathcal{X} & & \end{array} \tag{1}$$

We will refer to this diagram often, and to do so give it the somewhat non-descriptive, but in our context wholly unambiguous, name 'the standard diagram.'

Traditionally, $\pi_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$, defined by mapping $(x, y) \mapsto x$ (read: $\pi_{\mathcal{X}}(x, y) := x$), is taken to be "easy, efficient, or cheap" to evaluate or sample while $\pi_{\mathcal{Y}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$, defined by mapping $(x, y) \mapsto y$, is computationally expensive, expensive otherwise, difficult for other reasons, or altogether infeasible. The original space $\mathcal{X} \times \mathcal{Y}$ is itself inaccessible, except for some *labeled* data $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \mathcal{X} \times \mathcal{Y}$ which provides a proxy (and incomplete!) illustration of what $\mathcal{X} \times \mathcal{Y}$ looks like. The map $\tilde{y} : \mathcal{X} \dashrightarrow \mathcal{Y}$ is a critter we'd like to construct from data S so that both $\tilde{y}(x) \approx y$ for $(x, y) \in S$ *and* for $(x, y) \in \mathcal{X} \times \mathcal{Y}$. What "$\approx$" means, how to construct $\tilde{y}$, conditions on S which are needed to make this problem feasible, etc. are all aspects of the supervised machine learning problem which we will explore in this course.

As an example, suppose $\mathcal{X} = \mathbb{R}$ denotes credit score and $\mathcal{Y} = \{0, 1\}$ loan repayment (say '1' corresponds to repayment of loan, '0' to default). Then a *point* $(x, y) \in \mathbb{R}$ represents data corresponding to a loan whose account holder has credit score x and for which the loan was either paid in full ($y = 1$) or not ($y = 0$). The reason we say $\pi_{\mathcal{X}}$ is "easy" to sample is that you may ask any person what their credit score is (more realistically: as creditor, you would see this information *at the time of application*), while loan repayment information (the "label") would not be observed until potentially many years later when the loan is finally repaid or defaults.

It is worth noting, and perhaps lingering upon the observation, that the "input-output" relation (x, y) is not necessarily functional, i.e. for two points (x, y), $(x', y') \in \mathcal{X} \times \mathcal{Y}$, $x = x'$ does not imply that $y = y'$. The stand-in for determinism is probability, i.e. we will presume that there is some joint probability measure $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ on $\mathcal{X} \times \mathcal{Y}$, and e.g. that $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = f(x)|x) \neq 1$.

1.2.3 Themes

Generalization Given model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$, how well does $\tilde{y}$ match the (finite) data we have $S \subsetneq \mathcal{X} \times \mathcal{Y}$—i.e. $\tilde{y}(x) \approx y$ for $(x, y) \in S$—*and* the data we don't have, $(x, y) \in \mathcal{X} \times \mathcal{Y}$?

Dimensionality Computation in high dimensions becomes harder, in part because computation is more expensive, and because there are more "corners" for data to hide in (which exacerbates the computational problem). The geometry of high dimensionality will be a recurring theme; for now, we simply observe sources of dimensionality:

1. The data "set" itself $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \mathcal{X} \times \mathcal{Y}$. Properly speaking, this data will be presumed to be sampled $(x_i, y_i) \sim_{iid} \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ which *means* that the data "set" is a *point* in the space $(\mathcal{X} \times \mathcal{Y})^m$. A reasonable question to ask, then, is: what is the measure $\mathbb{P}_{(\mathcal{X}\times\mathcal{Y})^m}$? Independence tells us that it is $\prod_{j=1}^{m} \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$. More on this later.

2. Size of space itself. This could include high dimensionality of $\mathcal{X}$ and/or $\mathcal{Y}$. Examples abound of high dimensional input data: numerous columned tabular data, imagery data, audio data, video data.

3. Parameter space for model $\tilde{y}$. In the case of linear regression, a model $\tilde{y}(x) = \sum_{j=0}^{n} a_j x^j$ may have arbitrarily large degree n. Or a fully connected neural network with many nodes and many layers. And so on. More generally, the space of functions $\mathcal{Y}^{\mathcal{X}} := \{\tilde{y} : \mathcal{X} \to \mathcal{Y}\}$ is even bigger.

Trade-offs Assumptions must be made and compromises allowed for in order to gain tractability in the learning problem. There is no universal solution ("no free lunch," and as you may imagine, there's a theorem for that) and formulating the setup to address one challenge may introduce other ones elsewhere (ML can sometimes feel like one giant game of whackamole).

The famed bias-variance trade-off is one example: a high complexity model may well represent the data S, which in one sense is good, but in another is bad if said model represents data *too well*, i.e. at the exclusion of modeling 'from where the data comes.'

1.3 First Programming Assignment

You may find the first programming assignment under pa0 in git repo, and starter code in the git. I suggest you follow the tutorial at Real Python https://realpython.com/account/login/?next=/logistic-regression-python/[1] which shows you how to spin up a logistic regression model using sklearn. Scikit-Learn (also known as sklearn) is an open source ML library for python, and contains functionality for constructing numerous models. This is perhaps the only time in this course you will be asked to use this library, and if you have a preferred alternative library, you are more than welcome to use it for this assignment.

The purpose of the assignment is threefold:

1. Gain initial exposure to the *structure* of machine learning code, including object oriented programming and the typical methods included.
2. Shake off any residual rust using Python.
3. To gain deeper appreciation for the *aim* of building model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ as in diagram (2), and metrics that illustrate success.

Recall the Standard Diagram

$$\begin{array}{ccc} \mathcal{X} \times \mathcal{Y} & \xrightarrow{\pi_{\mathcal{Y}}} & \mathcal{Y} \\ \downarrow{\scriptstyle \pi_{\mathcal{X}}} & \nearrow_{\tilde{y}} & \\ \mathcal{X} & & \end{array} \tag{2}$$

This diagram provides formalism for talking about "approximating" y with model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ when $y \neq y(x)$ is not necessarily functionally determined by x.

Consider a concrete example to illustrate the problem: suppose that $\mathcal{X} = \mathbb{R}$ denotes credit score and $\mathcal{Y} = \{0, 1\}$ denotes repayment on loan, 1 denotes full repayment and 0 denotes default. A data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ corresponds to a credit score-loan repayment pair, and in real life a loan account would have these attributes associated with it, i.e. the debtor would have some credit score and their loan will (eventually) be repaid or not. (Note that "eventuality" is what, in this case, makes $\pi_{\mathcal{Y}}$ hard or expensive to evaluate.) It is possible for two different loans, belonging to two different people, to agree on credit score but disagree on outcome $y \in \mathcal{Y}$. In fact, we will likely observe both outcomes $y = 0$ and $y = 1$ associated to *any* credit score. Presumably, there should be some relation between the relative *counts* of $\#y = 1$ and credit score; in other words, one might suppose that lower credit scores correspond to accounts which in actual fact get repaid less frequently than those with high credit scores. We need mathematical language to describe and work with this phenomenon. The language is probability.

Thus, we suppose that there is some joint probability measure $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ on $\mathcal{X} \times \mathcal{Y}$. One way of contextualizing supervised machine learning is as a study of probability on the standard diagram. In this lecture, we will review probability, give intuition for probability as measure, as well as notation $\mathbb{P}_{\mathcal{X}} = \int d\mathbb{P}_{\mathcal{X}}(x)$, and define expectation $\mathbb{E}(f)$ of a random variable $f : \mathcal{X} \to \mathbb{R}$ as a *Lebesgue* integral $\mathbb{E}(f) := \int_{\mathcal{X}} f(x) d\mathbb{P}_{\mathcal{X}}(x)$. We kick this lecture off by emphasizing that probability has nothing to do with randomness... yet. When we return to admitting randomness into our lexicon, it will be as a *result*. For the moment, we forget any association between probability and chance, stochasticity, randomness, or any other (for now) anathema word affiliated with the notion of uncertainty.

[1] You may need to sign up in order to view this content, but it is not behind a paywall.

1.4 Intuition for Measure from Calculus

We start reviewing integration in calculus to preview notation for measure. One interpretation of the integral is as "area under the curve." Another (what amounts to similar) is as measure. And one interpretation of dQ is as an infinitesimal Q element, whatever Q is. Another is as an indicator of what kind of measurement we are taking. For example, length ℓ is $\int d\ell$, area a is $\int da$, volume v is $\int dv$, and so on. Calculus gives us a way to *compute* $\int dQ$, but here we instead suppose prior understanding of the basic unit of measurement Q and are defining the *notation* $\int dQ$ in terms of what we *know* already, namely Q.

Example: length $\ell = \int_{[a,b]} d\ell(x)$, where
$\ell(x)$ is a type of measure,
$[a, b]$ is an object we are measuring.

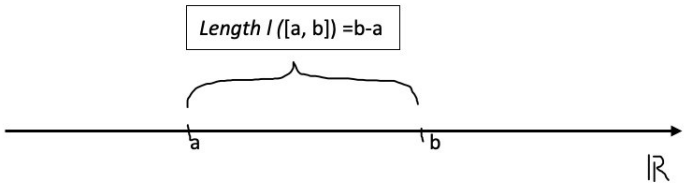

Figure 1: Example of length measure

Example: Area $\alpha(R) = \int_R d\alpha$, where we don't know what $d\alpha$ is but we define R.

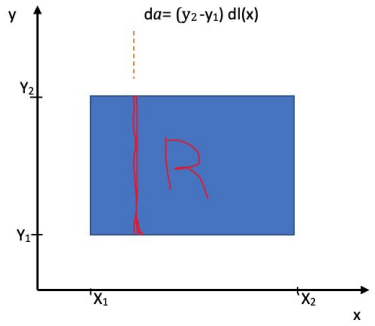

Figure 2: Example of Area measure

$$\mathbb{R} = [x_1, x_2] \times [y_1, y_2]$$

$$\mathbb{R} = (x_2 - x_1)(y_2 - y_1) = \ell([x_1, x_2])\ell([y_1, y_2])$$

$a(R) = \int_R da$, where
$a(R)$ is the area,
da is a unit of the area

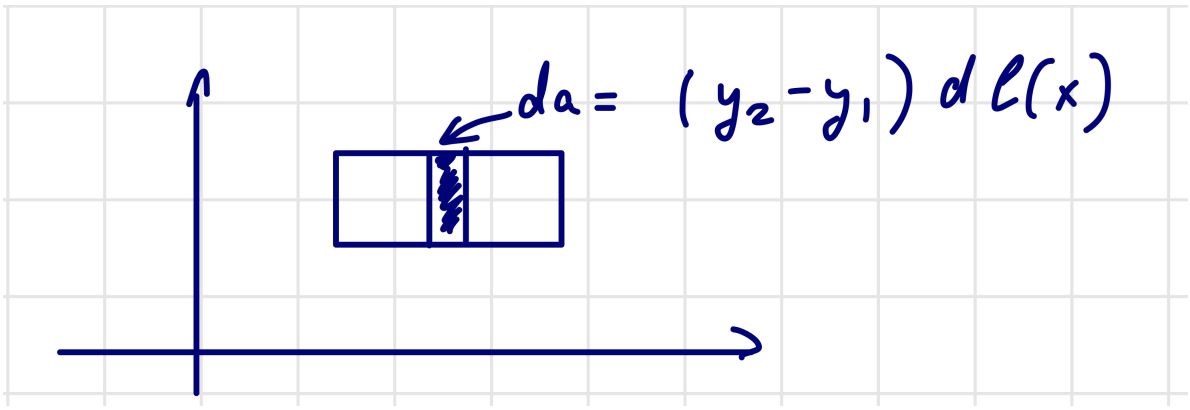

Figure 3: Area

Similarly,we can define $\mathbb{R}$ for volume **??** or curve
$\int_{curve} d\gamma = \gamma(curve)$
We can express this integral in terms of $X - Y$ coordinates.

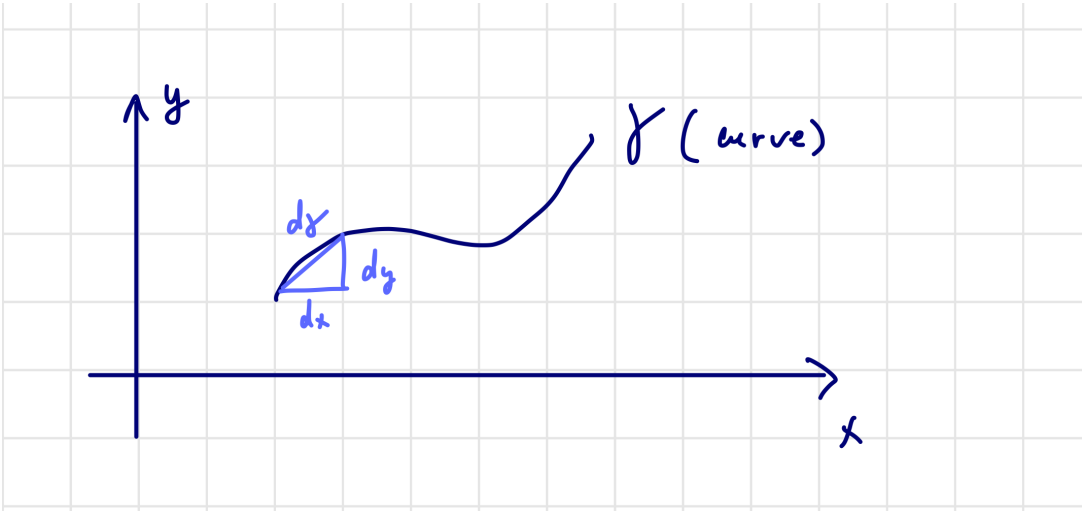

Figure 4: Curve in XY coordinates

Therefore, by Pythagorean theorem
$d\gamma^2 = dx^2 + dy^2 = dx^2(1 + (\frac{dy}{dx})^2)$
Thus, $d\gamma = \sqrt{1 + (\frac{dy}{dx}^2)}dx$
Therefore,
$\gamma(curve) = \int_{curve} \sqrt{1 + (\frac{dy}{dx})^2}dx$

1.5 The Probability Measure

Now we formalize what we mean by probability measure. The first point to make is that it is a measure.

Definition 1.1. Let $\mathcal{X}$ be a set. We define a *probability measure* $\mathbb{P}_{\mathcal{X}} : (\text{Subsets of } \mathcal{X}) \to [0,1]$ on $\mathcal{X}$ to be a map from (a subset of) the power set of $\mathcal{X}$ to the closed interval $[0,1]$ satisfying the following two properties:

1. $\mathbb{P}_{\mathcal{X}}(\mathcal{X}) = 1$, and
2. $\mathbb{P}_{\mathcal{X}}\left(\bigsqcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mathbb{P}_{\mathcal{X}}(A_j)$ where $\bigsqcup_{j=1}^{\infty} A_j$ denotes disjoint union, i.e. as a set $\bigsqcup_{j=1}^{\infty} A_j = \bigcup_{j=1}^{\infty} A_j$ and $A_i \cap A_j = \varnothing$ for $i \neq j$.

Recall that the power set $2^{\mathcal{X}}$ of a set $\mathcal{X}$ is defined to be the set of all subsets of $\mathcal{X}$, including $\varnothing$ and $\mathcal{X}$ itself. For example, when $\mathcal{X} = \{1,2,3\}$,

$$2^{\mathcal{X}} = \{\varnothing, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \mathcal{X}\}.$$

When $\mathcal{X} = \mathbb{R}$, the power set includes any reasonable combination (e.g. unions) of intervals (a,b) or $[a,b]$, but also many many more (see Cantor Set for a fun, but not particularly relevant, excursion into the wonders of measure theory).

The second condition is called 'countable additivity' (informally: a conservation of stuff principle) and represents the intuitive idea that if you slice and dice an object for measurement, measure each constituent piece without double counting, and add your results, you'll end up with the same result as if you just measured the original unadulterated tamale.

You should check that countable additivity implies *finite* additivity $\mathbb{P}_{\mathcal{X}}\left(\sum_{j=1}^{n} A_j\right) = \sum_{j=1}^{n} \mathbb{P}_{\mathcal{X}}(A_j)$. You will need the fact that $\mathbb{P}_{\mathcal{X}}(\varnothing) = 0$, which itself is implied by conditions 1. and 2. Indeed, $\mathcal{X} = \mathcal{X} \sqcup \bigsqcup_{j=2}^{\infty} \varnothing$.

Remark 1.1. We are being fairly blase about the *domain* "subsets of $\mathcal{X}$" of $\mathbb{P}_{\mathcal{X}}$. Perhaps we shouldn't be. Much of the architecture constructing measure depends on the fact that some structure must be placed on the set of subsets of $\mathcal{X}$ which are suitable for measurement; in particular, that such a set comprises a so-called σ-algebra, is not necessarily (and in many cases in fact is not) the entire power set $2^{\mathcal{X}} = \{A \subset \mathcal{X}\}$, and so on. We pay lip-service to this nuance, but fret little over the possibility that we will accidentally run across both a measure $\mathbb{P}_{\mathcal{X}}$ and subset $A \subset \mathcal{X}$ for which $\mathbb{P}_{\mathcal{X}}(A)$ does not make sense (read: which $\mathbb{P}_{\mathcal{X}}$ is "incapable" of measuring). One must try very hard—you might find such a question on a measure theory qualifying exam—to come up with an example. Therefore you may reasonably suppose that any set you'd come across in real life is in fact measurable. Still, know *that* there is a potential problem: if you can construct a non-measurable set, then you can cut an apple into finitely many pieces and reassemble those finitely many pieces into *two* apples of the same size (see Banach Tarski for more information). In other words, weird things can happen with the things that aren't measurable.

Definition 1.2. We define a *probability space* to be a pair $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ where $\mathcal{X}$ is a set and $\mathbb{P}_{\mathcal{X}} : (\text{Subsets of } \mathcal{X}) \to [0,1]$ is a probability measure (c.f. definition 1.1).

I won't vehemently object if you use residual terminology to describe things, e.g. measurable subsets $A \subset \mathcal{X}$ are often called 'events' and individual points $x \in \mathcal{X}$ called 'outcomes.' My only objection is if you associate this language with the notion of randomness. Remember for now: probability is not about randomness!

Example 1 Let $\left(\mathcal{X} = [0,1], \mathbb{P}_X([a,b]) := b - a\right)$ for $0 \leq a \leq b \leq 1$. We define notation $\int_{[a,b]} d\mathbb{P}(x) = \mathbb{P}([a,b])$.

Example 2 Let $\left(\mathcal{X} = \mathbb{R}, \mathbb{P}_X([a,b]) := \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\right)$, the so-called normal distribution with zero mean and unit variance. Observe that we use a Riemann integral to *compute* or give the rule for realizing the probability measure. The integrand $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ of the Riemann integral is called a *probability density function*. The integral $\int_{[a,b]} d\mathbb{P}_{\mathcal{X}}(x)$, by contrast, is not a Riemann integral; it is *defined* by the measure $\mathbb{P}_{\mathcal{X}}([a,b])$. (When you ask: but what is the measure?, we gave the rule for how to calculate it!)

Example 3 Let $\mathcal{Y} = [0,1]$, $\mathbb{P}_{\mathcal{Y}}([x_1,x_2]) \times [y_1,y_2]) = (x_2 - x_1) \cdot (y_2 - y_1)$. This example is a preliminary look into independence as $\mathbb{P}_{\mathcal{X}}([x_1,x_2]) = x_2 - x_1$, $\mathbb{P}_{\mathcal{Y}}([y_1,y_2]) = y_2 - y_1$ and $\mathbb{P}_{\mathcal{X}}([x_1,x_2] \times [y_1,y_2]) = \mathbb{P}_{\mathcal{X}}([x_1,x_2])\mathbb{P}_{\mathcal{Y}}([y_1,y_2])$ We will return to this example when we discuss independence, and will want to situate as a notion which is at home in high dimension.

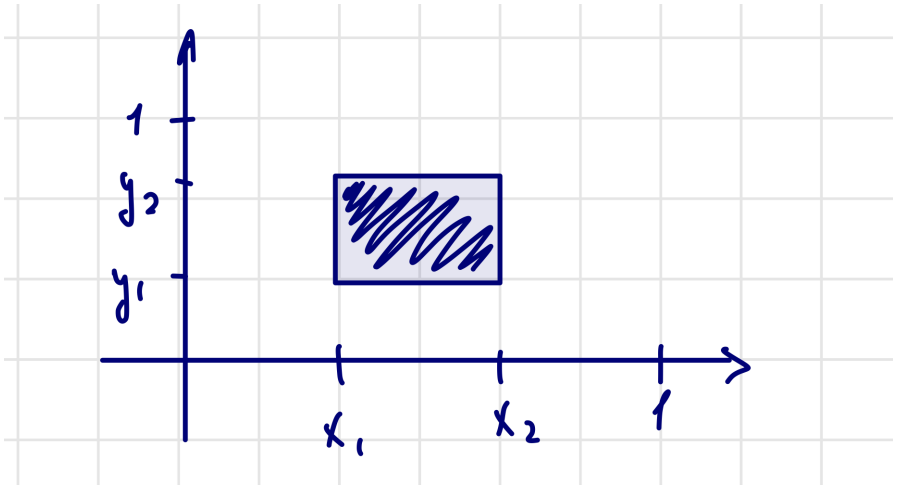

Figure 5: The area measure

Independence is a high dimensional phenomenon as it is clear that we can generalize the picture in 5 to n dimensions.

Example 4 : Bernouli Random Variable (flipping a fair or unfair coin): Let $\mathcal{X} = \{0,1\}$ where $\mathbb{P}_{\mathcal{X}}(\{1\}) = p$ and $\mathbb{P}_{\mathcal{X}}(\{0\}) = 1 - p$. This is a probability space with two outcomes that clearly satisfies the axioms stated earlier since the events or subsets of the space X, $\{0\}$ and $\{1\}$ are disjoint and the sum of their probabilities is 1:

$$\mathbb{P}_X(1) + \mathbb{P}_{\mathcal{X}}(0) = p + (1-p) = 1 = \mathbb{P}_X(\{0\} \sqcup \{1\})$$

1.6 Random Variables

Definition. Recall probability space $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ consists of a set $\mathcal{X}$ and a measure $\mathbb{P}_{\mathcal{X}}$ (definition 1.2). We noted at the beginning of lecture that probability is properly a theory of measure, which *really* means that it is a theory of integration. Which deals with numbers (values). So far, we've said nothing about the nature of the set $\mathcal{X}$. We don't need to. But we do need a way to ascribe values to outcomes $x \in \mathcal{X}$.

Definition 1.3. Let $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ be a probability space. We define a *random variable* to be a *function* $f : \mathcal{X} \to \mathbb{R}$, whose codomain is $\mathbb{R}$.

We observe that such a function induces a measure $\mathbb{P}_{\mathbb{R}}$ on $\mathbb{R}$, defined by

$$\mathbb{P}_{\mathbb{R}}([a,b]) := \mathbb{P}_{\mathcal{X}}(f^{-1}([a,b])) = \mathbb{P}_{\mathcal{X}}(\{x \in \mathcal{X} : f(x) \in [a,b]\}). \tag{3}$$

One should check that this defines an honest probability measure on $\mathbb{R}$.

Remark 1.2. It is worth noting that there are conditions on the map $f : \mathcal{X} \to \mathbb{R}$ needed in order to ensure that the induced measure $\mathbb{P}_{\mathbb{R}}$ is well-defined, i.e. actually is a (probability) measure. The condition is called 'measurability,' i.e. we require f to be a *measurable* function (which really just means: f is such that the induced measure is a measure—this isn't circular!, it all comes down to saying, you cannot with impunity claim that any function whatsoever will induce a measure). Just as with the domain of $\mathbb{P}_{\mathcal{X}}$, where we suppose with little guilt that "all subsets" we encounter are measurable, we will also suppose that the functions we come across in practice are measurable. Again, there is nuance to be appreciated, but the supposition we make will very unlikely harm any of our day to day calculations. (For the ultra-curious, the condition of measurability stipulates that any potentially measurable set in $\mathbb{R}$ has preimage (by f^{-1}) which *is* measurable in $\mathcal{X}$.)

Remark 1.3. We noted that $f : \mathcal{X} \to \mathbb{R}$ induces a measure. In fact, this holds more generally for any (measurable) map $f : \mathcal{X} \to \mathcal{Y}$, $\mathbb{P}_{\mathcal{Y}}(B) := \mathbb{P}_{\mathcal{X}}(f^{-1}(B))$.

Definition 1.4. Let $f : \mathcal{X} \to \mathbb{R}$ be a random variable (definition 1.3). We define *expectation of* f, denoted $\mathbb{E}(f)$, to be

$$\mathbb{E}(f) := \int_{\mathcal{X}} f(x) d\mathbb{P}_{\mathcal{X}}(x). \tag{4}$$

Equation (4) defines expectation, but we have some odd sort of critter we've not seen before on the right hand side. We must define *it*.

Definition 1.5. Let $f : \mathcal{X} \to \mathbb{R}$ be a *simple* random variable, i.e. taking *finitely* many values $a_1, \ldots, a_k$. Then we define the *Lebesgue* integral of f, denoted $\int_{\mathcal{X}} f(x) d\mathbb{P}_{\mathcal{X}}(x)$, to be

$$\int_{\mathcal{X}} f(x) d\mathbb{P}_{\mathcal{X}}(x) := \sum_{j=1}^{k} a_j \dot{\mathbb{P}}_{\mathcal{X}}(f = a_j) = \sum_{j=1}^{k} a_j \mathbb{P}_{\mathcal{X}}\left(\{x \in \mathcal{X} : f(x) = a_j\}\right).$$

One may recognize this definition as the *expectation of a discrete random variable*. Indeed, that is exactly what it is. It is the Lebesgue integral! The Lebesgue integral extends to continuous random variables, but it requires some abstraction. For next time. In the meantime, note that for continuous random variable $f : \mathcal{X} \to \mathbb{R}$ (not necessarily taking finitely many values), we can *approximate* $\mathbb{E}(f)$ using the Lebesuge integral of a simple random variable (expectation of discrete r.v.), see fig. 6.

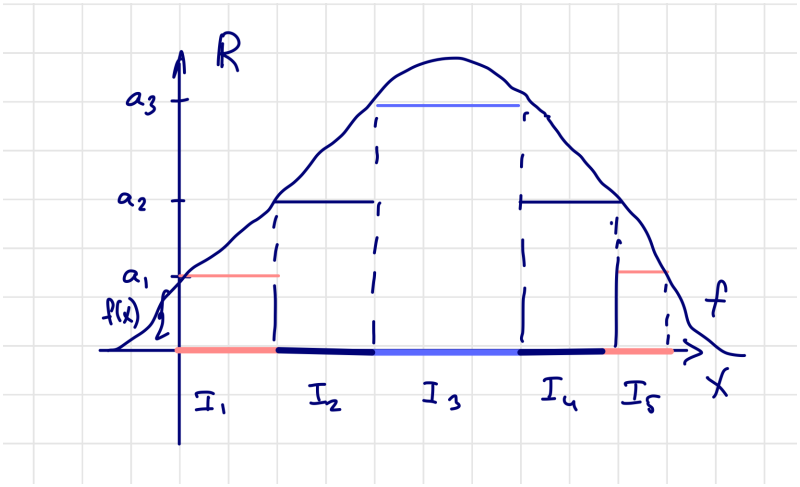

Figure 6: Piecewise defined function f

Remark 1.4. In **??**, we will continue the definition of Lebesgue integral for continuous random variable. While there are theoretical reasons for insisting on the use of this abstraction, ours are more practically oriented. In fact, one may (very) often compute expectation *using* instead a Riemann integral. For example, the expectation of a mean 0 unit variance normally distributed random variable

is $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-x^2/2}dx$ (which of course is zero). In other words, if push comes to shove and you're asked to *compute* the expectation for a continuous random variable, very often you'll have a density function in your pocket and can just multiply it by the random variable, and integrate, business as usual. We introduce Lebesgue integration for two reasons:

1. To articulate the fact that the extension from discrete to continuous random variables may *seem* confusing, and that is due in part to the nauseating head-spinning move from Lebesgue to who knows (but usually Riemann) integration without even lip-service paid to the fact that expectation of discrete r.v.s itself is a non-trivial extension of our conceptual apparatus.

2. Ease of notation: $d\mathbb{P}_{\mathcal{X}}$ always makes sense and makes immediately obvious what our measure is. In other words, I don't always want to say: suppose that the density of a probability space exists, and anyway it might not and that doesn't matter!, for $\int_A d\mathbb{P}_{\mathcal{X}}(x)$ makes sense regardless of whether we can integrate (Riemann-wise) as $\int \varphi(x)dx$ (for density $\varphi(x)$). Related: the notation $d\mathbb{P}_{\mathcal{X}}$ collapses the distinction between continuous and discrete random variables. The distinction, in my view, is convoluted and confusing, especially when we have mixed discrete-continuous nonsense going on (e.g. in the case of binary classificaiton $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$). Of course, successfully working with the collapse requires an comfort with the abstraction, and that may by itself be initially confusing as well.

Remark 1.5. Observe also that $\mathbb{E}(f)$ is somewhat uninformative, in a way that $\int_{\mathcal{X}} f(x)d\mathbb{P}_{\mathcal{X}}(x)$ is not. At the moment the added notational baggage of the latter may seem inconvenient, but once we start squinting at joint probability spaces $\mathcal{X} \times \mathcal{Y}$, turning them upside down, and so on, it will be imperative to be clear on how we are integrating. For example, we will see

$$\int_{\mathcal{X}\times\mathcal{Y}} f(x,y)d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x,y) = \int_{\mathcal{X}}\int_{\mathcal{Y}|\mathcal{X}} f(x,y)d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)d\mathbb{P}_{\mathcal{X}}(x).$$

This is sortof an extension on the notational point above. I will be relying on this notation aggressively, so it's crucial to anticipate eventually becoming comfortable with standard manipulations, and recalling (any time there is lingering confusion) that $d\mathbb{P}_{\mathcal{X}}$ is defined, not in isolation, but as a package $\int_A d\mathbb{P}_{\mathcal{X}} := \mathbb{P}_{\mathcal{X}}(A)$.