Machine Learning I

Lecture 2: Foundations

Nathaniel Bade

Northeastern University Department of Mathematics

Table of contents

1. Elements of Probability
2. Continuous Probability Distributions
3. Continuous Joint Distributions
4. Expectation Value
5. A Formal Model for Statistical Learning
6. Empirical Risk Minimization
7. Proving Bounds Using Realizability and I.I.D.

Elements of Probability

Elements of Probability:

Probability is the study of **random variables**. A random variable is a probabilistic outcome, such as rolling a number on a dice, tossing a coin and getting heads, or the blood glucose level of a random subject in an experiment.

There are three elements that make up probabilistic models:

Sample Space.

Set of events.

Probability distribution.

Elements of Probability: Sample Space

A **random variable** parameterizes a **sample space** S. This space may be discrete or continuous.

For example,

> **Coin**: $S = \{H, T\}$
>
> **Dice**: $S = \{1, 2, 3, 4, 5, 6\}$.
>
> **Blood Glucose**: $S = (0, \infty) = \mathbb{R}_+$.

Each outcome $x_i \in S$ is a complete description of the state of the real world at the end of the experiment.

Elements of Probability: Event Space

For a given sample space the **event space** (or **set of events**) $\mathcal{F}$ is a set whose elements are subsets of S. Each subset $A \in \mathcal{F}$ is an **event**. The event space is identified with the powerset of S: $\mathcal{F} = \text{Pow}(S)$.

For example, in dice rolling an event could be $A = \{1, 3, 5\} \in \mathcal{F}$, or rolling an odd number. All such events are contained in $\mathcal{F}$.

In the blood glucose example, a fasting blood glucose level of above 100 but below 125 mg/dL is considered prediabetic. If we measure the blood glucose of a random test subject, one **event** we could be looking for is prediabetes. Mathematically, this event would be the half open interval $A = (100, 125]$ in the sample space $\mathbb{R}_+$.

Elements of Probability: Measure

The probability **measure** or **distribution** is a function $\mathcal{D} : \mathcal{F} \to \mathbb{R}$ that satisfies the following properties:

(a) $\mathcal{D}(A) \geq 0$ for all events A.

(b) $\mathcal{D}(S) = 1$

(c) If A_i are disjoint events, then $\mathcal{D}\left(\bigcup_i A_i\right) = \sum_i \mathcal{D}(A_i)$

It is common to denote the probability of an event A given some density by $\mathbb{P}_{\mathcal{D}}(A)$. If the density is unambigious, the probability of an event is also denoted $\mathbb{P}(A)$.

Discrete Distributions

A **discrete distribution** assigns a probability to each value in a discrete sample space. For example, if X is an unfair dice, the sample space contains events $\{X = 1, \ldots, X = 6\}$ and the distribution might look like

$\mathcal{D}$:

x	1	2	3	4	5	6
$\mathbb{P}(X = x)$	.1	.1	.1	.5	.1	.1

Then $\mathbb{P}(X = 3) = .1$ and $\sum_x \mathbb{P}(X = x) = 1$.

Joint Distributions

Unfair Dice:

$\mathcal{D}$:

x	1	2	3	4	5	6
$\mathbb{P}(X = x)$	.1	.1	.1	.5	.1	.1

If we have *n* unfair dice, the probability distribution over the space of all outcome is called the **joint distribution** $\mathcal{D}^n$.

If all of the variables are uncorrelated,
$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$.

Question: What is $\mathbb{P}_{\mathcal{D}^4}(4444)$?

Ans: $(.5)^4 = 0.0625$.

Joint Distributions

Events for joint distributions may or may not be defined disjointly. For example, if $A = \{2, 3, 4\}$ then the event $\mathbb{P}(X \in A, Y = 2)$ contains the outcomes

$$\mathbb{P}(X = x, Y = y)$$

y/x	1	2	3	4	5	6
1	.01	.01	.01	.05	.01	.01
2	.01	.01	.01	.05	.01	.01
3	.01	.01	.01	.05	.01	.01
4	.05	.05	.05	.25	.05	.05
5	.01	.01	.01	.05	.01	.01
6	.01	.01	.01	.05	.01	.01

The probability of $\mathbb{P}(X \in A, Y = 2) = .07$.

Joint Distributions

On the other hand, let $A = \{(x, y) : 2 < x + y \leq 6\}$.

$$\mathbb{P}(X = x, Y = y)$$

y/x	1	2	3	4	5	6
1	.01	.01	.01	.05	.01	.01
2	.01	.01	.01	.05	.01	.01
3	.01	.01	.01	.05	.01	.01
4	.05	.05	.05	.25	.05	.05
5	.01	.01	.01	.05	.01	.01
6	.01	.01	.01	.05	.01	.01

Question: What is $\mathbb{P}(A)$?

Joint Distributions

On the other hand, let $A = \{(x, y) : 2 < x + y \leq 6\}$.

$$\mathbb{P}(X = x, Y = y)$$

y/x	1	2	3	4	5	6
1	01	.01	.01	.05	.01	.01
2	.01	.01	.01	.05	.01	.01
3	.01	.01	.01	.05	.01	.01
4	.05	.05	.05	.25	.05	.05
5	.01	.01	.01	.05	.01	.01
6	.01	.01	.01	.05	.01	.01

The probability of $P(A) = \sum_{(x,y)\in A} P(X = x, Y = y) = .3$.

Conditional Distributions

A **conditional distribution** is the distribution of a random variable given an observation about another random variable. For example

$$\mathbb{P}(X = x | Y = y)$$

is the probability that $X = x$ given that $Y = y$.

In general, the conditional probability of any event A given that the event B is observed is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} .$$

Two events are called **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$, or equivalently if $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Conditional Distributions

Conditional probability:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}\,.$$

Question: In basketball, it is claimed that in a shootout situation the success rate of successive shots is independent. Say X_1 and X_2 are shots 1 and 2. For a specific player, the probability of getting $\mathbb{P}(X_1 = \text{Score}) = .49$. What must $\mathbb{P}(X_1 = \text{Score} \cap X_2 = \text{Score})$ be if the shots are actually independent?

(a) .49

(b) .51

(c) $\mathbb{P}(X_2 = \text{Score})$

(d) $\mathbb{P}(X_1 = \text{Score})\mathbb{P}(X_2 = \text{Score})$

Conditional Distributions

Conditional probability:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} .$$

Question: In basketball, it is claimed that in a shootout situation the success rate of successive shots is independent. Say X_1 and X_2 are shots 1 and 2. For a specific player, the probability of getting $\mathbb{P}(X_1 = \text{Score}) = .49$. What must $\mathbb{P}(X_1 = \text{Score} \cap X_2 = \text{Score})$ be if the shots are actually independent?

(a) .49

(b) .51

(c) $\mathbb{P}(X_2 = \text{Score})$

(d) $\mathbb{P}(X_1 = \text{Score})\mathbb{P}(X_2 = \text{Score})$

Continuous Probability Distributions

Continuous Distributions

Random variables may also be continuous. Instead of defining a probability distribution element-wise, with all probabilities summing to 1, we often define a probability **density** $p(x)$ which integrates to 1.

For example, if the sample space for X is $\mathbb{R}$, then

$$\int_{-\infty}^{\infty} p(x)\, dx = 1\,.$$

Continuous Distributions

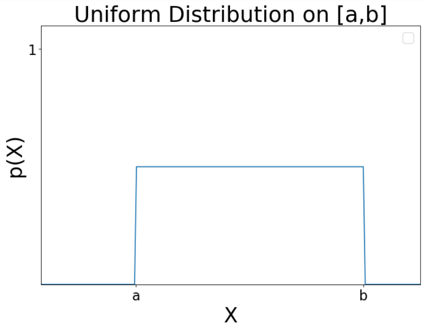
The probability density function is defined over S and is **not** the probability $X = x$. Instead, for any event $A \subset S$, we obtain the probability of A by integrating:

$$\mathbb{P}(A) = \int_A p(x) dx \,.$$

Therefore $p(x)$ defines a probability **distribution** via $\mathcal{D}(A) = \mathbb{P}(A)$.

Exercise: For a $S = \mathbb{R}$ and p a smooth probability density function, show that $\mathbb{P}(X = x) = 0$ for any fixed $x \in S$.

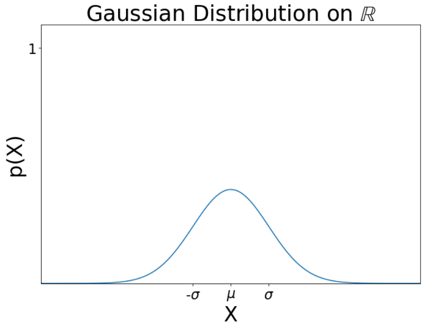
Common Distributions



The uniform distribution on $[a, b]$:

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

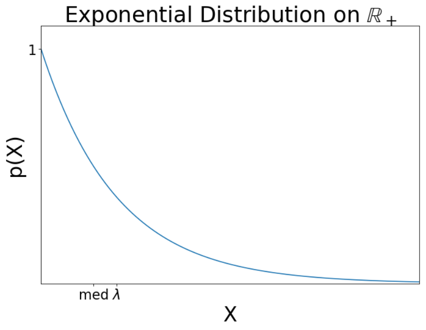
Common Distributions



The Gaussian distribution on $\mathbb{R}$:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{1}{2\sigma^2}(x-\mu)^2}$$

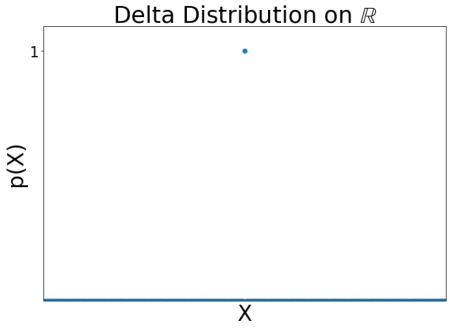
Common Distributions



The exponential distribution on $\mathbb{R}_+$:

$$p_\lambda(x) = \lambda e^{-\lambda x}$$

Common Distributions

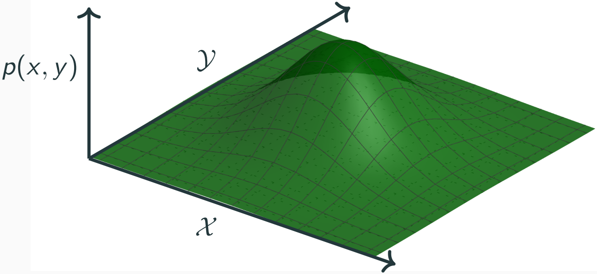


The delta distribution on $\mathbb{R}$ at a:

$$\delta_a(x) = \begin{cases} 1 & \text{if } x = a \\ 0 & \text{otherwise} \end{cases}$$

Continuous Joint Distributions

Joint Distributions

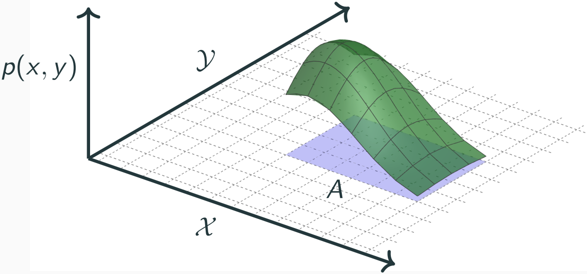


If X and Y are random variables parameterizing $\mathcal{X}$ and $\mathcal{Y}$ respectively, then a joint probability density

$$p_{X,Y} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

is just a probability density on $S = \mathcal{X} \times \mathcal{Y}$.

Joint Distributions

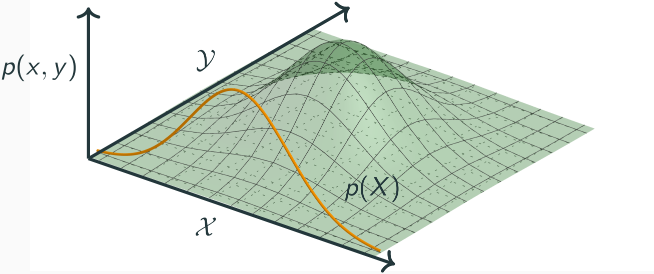


Integrating $p_{X,Y}(x, y)$ over events $A \subset \mathcal{X} \times \mathcal{Y}$ leads to a probability distribution

$$\mathcal{D}(A) = \int_A p(x, y)\, dx\, dy$$

where the probability of an event is $\mathbb{P}(A) = \mathcal{D}(A)$.

Marginal Probability

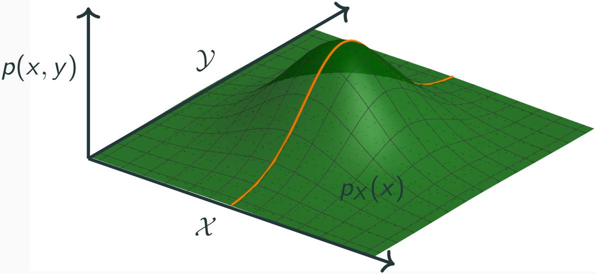


The **marginal** probability density is the density of one variable alone, obtained by integrating over all the contributions from the other variable:

$$p_X(x) = \int_{\mathcal{Y}} p(x, y) dy \,.$$

Note: it is not the projection onto $\mathcal{X}$.

Conditional Probability

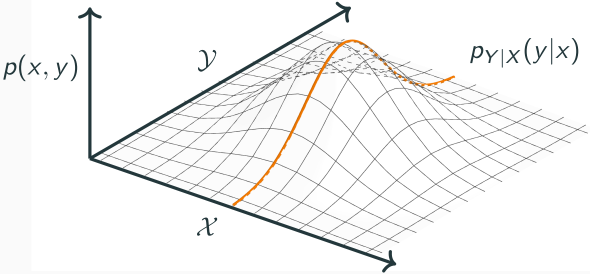


The **conditional** probability density allows us to find the distribution on Y when we know X must take a certain value. We define this conditional density as

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

Dividing by the marginal probability ensures that $p_{Y|X}(y|x)$ is in fact a distribution.

Conditional Probability

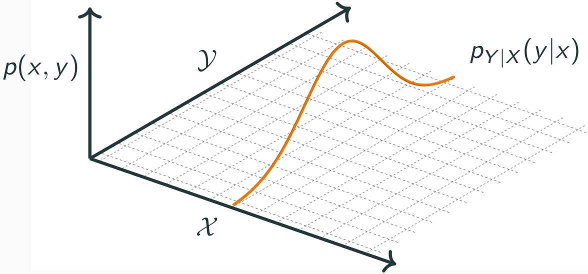


The **conditional** probability density allows us to find the distribution on Y when we know X must take a certain value. We define this conditional density as

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

Dividing by the marginal probability ensures that $p_{Y|X}(y|x)$ is in fact a distribution.

Conditional Probability

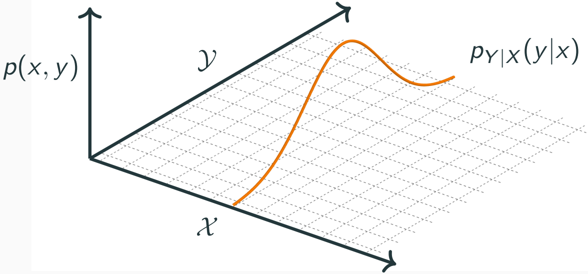


The **conditional** probability density allows us to find the distribution on Y when we know X must take a certain value. We define this conditional density as

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

Dividing by the marginal probability ensures that $p_{Y|X}(y|x)$ is in fact a distribution.

Conditional Probability



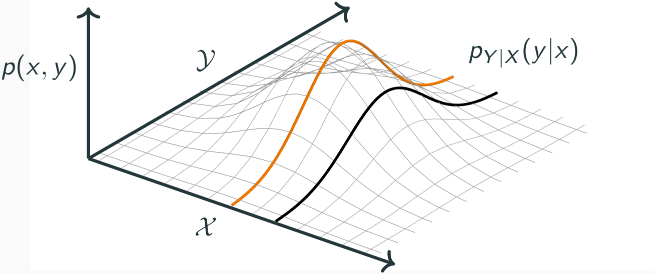
Random variables are **independent** if knowing about X tells us nothing about Y, that is

$$p_{Y|X}(y|x) = p_Y(y).$$

Then the density factorizes

$$p_{X,Y}(x, y) = p_X(x)p_{Y|X}(y|x) = p_X(x)p_Y(y)\,.$$

Conditional Probability



Random variables are **dependent** if $p_{Y|X}(y|x)$ changes with x.

Expectation Value

Expectation Value

If $f(X)$ is a function of a random variable, than $f(X)$ is itself a random variable. The **expectation value** of $f(X)$ is the weighted average

$$E[f(X)] = \sum_x f(x)\mathbb{P}(X = x) \qquad \text{Discrete}$$

or

$$E[f(X)] = \int_S f(x)p(x)dx \qquad \text{Continuous}$$

Conditional Expectation Value

The conditional expectation value is defined similarly:

$$E(f(X)\,|\,Y=y)=\int_{\mathcal{X}} f(x)p_{X|Y}(x|y)dx$$

Recalling the definition of the conditionalized density,

$$E(f(X)\,|\,Y=y)=\int_{\mathcal{X}} f(x)\frac{p_{X,Y}(x,y)}{p_Y(y)}dx$$

Conditional Expectation Value

Example: Let $S = \{(X, Y) : 0 < X < Y < \infty\}$ and define a joint density on S by

$$p(x, y) = e^{-y} \, .$$

We want to compute the mean of the conditional distribution

$$E(Y|X = x) = \int_{\mathcal{Y}} y \frac{p(x, y)}{p_X(x)} \, dy \, ,$$

Conditional Expectation Value

Example Cont: First first compute the marginal distribution

$$p_X(x) = \int_x^\infty p(x,y)\,dy = \int_x^\infty e^{-y}\,dy = e^{-x}\,.$$

Then

$$E(Y|X=x) = \int_x^\infty y\,\frac{p(x,y)}{p_X(x)}\,dy = \int_x^\infty y e^{x-y}\,dy\,.$$

Integrating out y gives

$$E(Y|X=x) = e^{x-y} + y e^{x-y}\Big|_x^\infty = 1 + x.$$

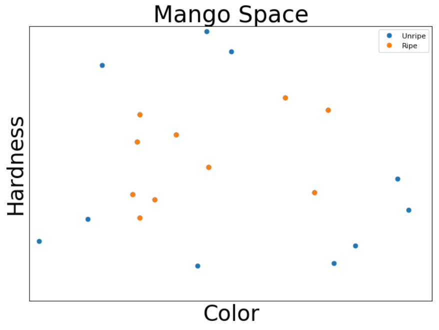
A Formal Model for Statistical Learning

Formal Model

We will begin to construct a formal model for machine learning using the **partially approximately correct** model of Shalev-Shwartz and Ben-David. The text is highly encouraged, but I will include in the slides everything necessary for the homework. We will however strive to use the notation of Hastie, Tibshirani and Friedman.

The goal of this lecture is to begin to provide a rigorous framework for statistical learning. We will first build our theory out in a rather simplified setting and then add layers of complexity.

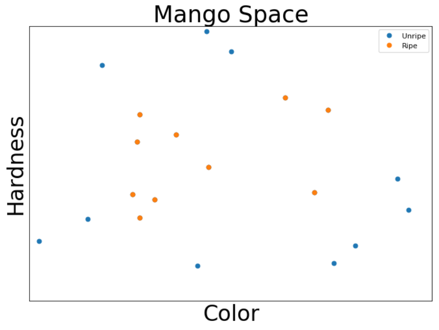
Formal Model: Example



In this lecture we will be considering supervised learning with a discrete label set. In supervised learning, a **learner** is an algorithm that will produce a function assigning a label to each element of some domain.

A guiding example: the perfect mango.

Formal Model: Example



Assume you have charted all the mangos you eat by color and hardness, recording which are perfectly ripe. A **learner** produce an algorithmic rule from this data set to label any mango as ripe or unripe.

Formal Model: Inputs and Outputs

In the basic setting, the learner has access to the following:

Learner Inputs:

Domain Set: A set $\mathcal{X}$ providing a domain for the input X.

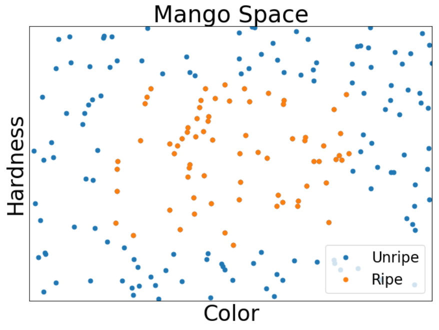
Label Set: A (discrete) set $\mathcal{Y}$ providing the range for the output Y.

Training Data: A sequence of N pairs $\mathcal{T} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ in $\mathcal{X} \times \mathcal{Y}$.

Learner Output:

The **learners output** will be a **prediction rule** $h : \mathcal{X} \to \mathcal{Y}$. This function is also called a **predictor** or **classifier**. We denote by $A(\mathcal{T}) = h_{\mathcal{T}}$ the **predictor** returned by an algorithm A run on the training set $\mathcal{T}$.

Formal Model: Distribution Dependence

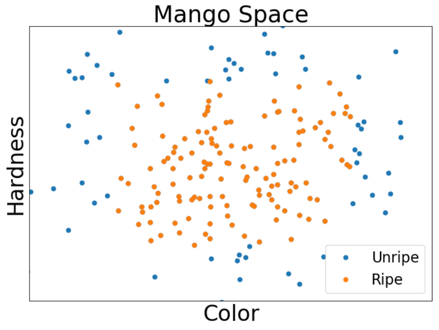


Data generation: We assume that the mangos we have tasted are generated by a probability distribution $\mathcal{D}$ on $\mathcal{X}$.

We also assume that there is a "correct" labeling $f : \mathcal{X} \rightarrow \mathcal{Y}$. This will be relaxed shortly.

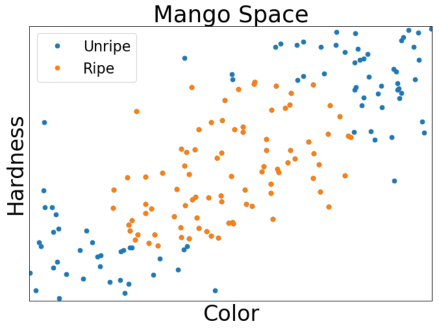
For example, if $f(x)$ labels as ripe all points inside the central square, a change in distribution will changed the observed training data: **Uniform**

Formal Model: Distribution Dependence



Data generation: We assume that the mangos we have tasted are generated by a probability distribution $\mathcal{D}$ on $\mathcal{X}$.
We also assume that there is a "correct" labeling $f : \mathcal{X} \rightarrow \mathcal{Y}$. This will be relaxed shortly.
For example, if $f(x)$ labels as ripe all points inside the central square, a change in distribution will changed the observed training data: **Gaussian**

Formal Model: Distribution Dependence



Data generation: We assume that the mangos we have tasted are generated by a probability distribution $\mathcal{D}$ on $\mathcal{X}$.

We also assume that there is a "correct" labeling $f : \mathcal{X} \to \mathcal{Y}$. This will be relaxed shortly.

For example, if $f(x)$ labels as ripe all points inside the central square, a change in distribution will changed the observed training data: **Other**

Empirical Risk Minimization

Empirical Risk Minimization

Given a training set $\mathcal{T}$, a learner must return a reasonable classifier. Without access to the distribution $\mathcal{D}$ or the correct labeling f, the true error is not available.

Define the training error to be

$$L_{\mathcal{T}}(h) := \frac{\left|\{i : h(x_i) \neq y_i\}\right|}{N}.$$

This is also called the **empirical error** or **empirical risk**. **Empirical risk minimazation** is the learning paradigm that states that this risk should be minimized.

Error

The **error of a classifier** h is the probability h does not predict the correct label for a point x drawn from $\mathcal{X}$ according to the distribution $\mathcal{D}$:

$$\textbf{Error}: \qquad L_{D,f}(h) := \mathbb{P}_{x\sim D}\big[h(x) \neq f(x)\big] .$$

We will also use the notation

$$\textbf{Error}: \qquad L_{D,f}(h) := \mathcal{D}\big(\,\{x : h(x) \neq f(x)\}\big) .$$

The error is the probability of randomly choosing an example x for which $f(x) \neq h(x)$.

This is also known as the **risk** or the **true error**.

Alternative Errors

Other notions of error: We would only use the error

$$L_{D,f}(h) := \mathbb{P}_{x \sim D}\left[h(x) \neq f(x)\right]$$

for categorical variables, and then for ones with no extra structure.

For quantitative variables, the error function is a generalization of RSS

$$\mathrm{Err}_{\mathcal{T}} = E\left[(h(x) - f(x))^2\right].$$

This error is useful for quantitative outputs or ordered discrete outputs.

Biased Empirical Risk Minimization

The **empirical risk minimization** (**ERM**) rule tells us we should find a predictor h that minimizes the training error

$$L_{\mathcal{T}}(h)\,.$$

If we let h be an arbitrary function, we immediately overfit, since h can take arbitrary values for each $x \in \mathcal{X}$.

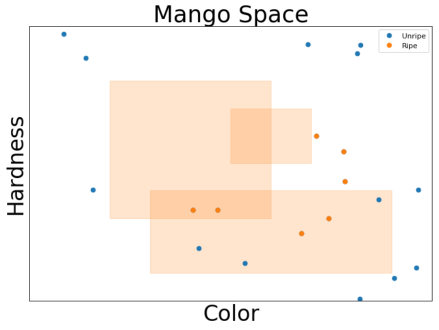
A common solution is to apply **ERM** only over a restricted function space called a **hypothesis class**.

For a hypothesis class $\mathcal{H}$, the $\textbf{ERM}_{\mathcal{H}}$ chooses a predictor from $\mathcal{H}$ with the lowest error over $\mathcal{T}$:

$$\textbf{ERM}_{\mathcal{H}}(\mathcal{T}) \in \operatorname{argmin}_{h \in \mathcal{H}} L_{\mathcal{T}}(h)\,.$$

Such a restriction is called an **inductive bias**.

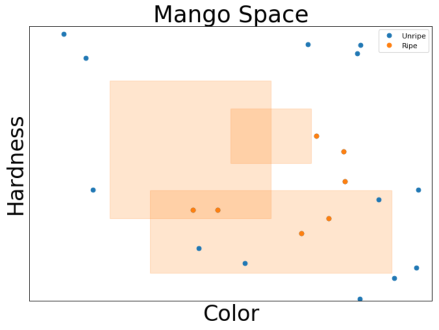
Biased Empirical Risk Minimization: Examples



The **indicator function** for any set $S \subseteq \mathcal{X}$ is

$$1_S(x) = \begin{cases} 1 & x \in S \\ 0 & x \notin S \end{cases},$$

Biased Empirical Risk Minimization: Examples

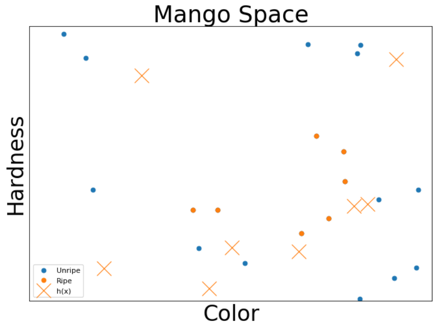


In the mango example, would could pick

$$\mathcal{H} = \{1_{[a,b]\times[c,d]} \,|\, a, b, c, d \in [0,1]\}$$

to be the set of indicator functions that is 1 on the interior of the square $[a, b] \times [c, d]$ and 0 otherwise. We will see that restricting to this class *cannot* overfit.

Biased Empirical Risk Minimization: Examples

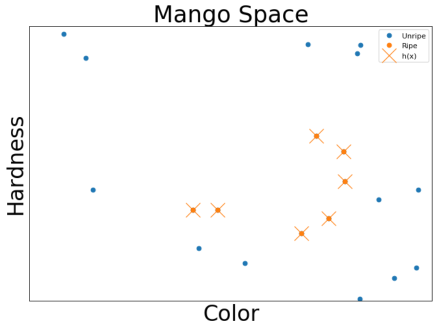


On the otherhand, we could pick the hypothesis class to be the set of functions that are 1 on some finite number of points

$$\mathcal{H} = \{1_{x_1,\ldots,x_n} \mid x_i \in [0,1], n \in \mathbb{Z}\}\,.$$

It is clear that restricting to this class will always allow overfitting.

Biased Empirical Risk Minimization: Examples



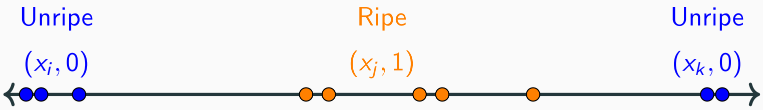
On the otherhand, we could pick the hypothesis class to be the set of function that is 1 on a finite number of points

$$\mathcal{H} = \{\delta_{x_1,\dots,x_n} \mid x_i \in [0,1], n \in \mathbb{Z}\}\,.$$

It is clear that restricting to this class will always allow overfitting.

Proving Bounds Using Realizability and I.I.D.

Mango Space



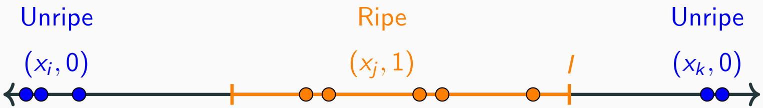
$$\mathcal{X} = \mathbb{R}$$

We will compute the number of data points need to fit the 1D mango example. The input space will be $\mathcal{X} = \mathbb{R}$ and the label space will be $\mathcal{Y} = \{0, 1\}$.

The hypothesis class is the class of indicator functions on intervals in $\mathbb{R}$

$$\mathcal{H} = \{1_I : I = [a, b]\}.$$

Mango Space



$$\mathcal{X} = \mathbb{R}$$

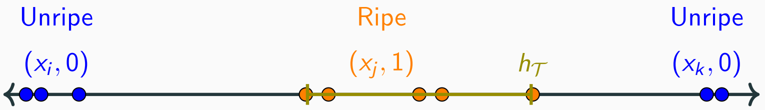
As a first example, we will compute the number of data points need to fit the 1D mango example. The input space will be $\mathcal{X} = \mathbb{R}$ and the label space will be $\mathcal{Y} = \{0, 1\}$.

The hypothesis class is the class of indicator functions on intervals in $\mathbb{R}$

$$\mathcal{H} = \{\delta_I \,:\, I = [a, b]\}\,.$$

The $\mathbf{ERM}_{\mathcal{H}}$ rule will be implemented by finding the smallest interval containing all data points labeled 1.

Mango Space



$$\mathcal{X} = \mathbb{R}$$

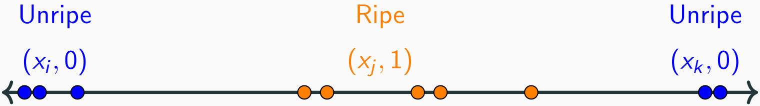
As a first example, we will compute the number of data points need to fit the 1D mango example. The input space will be $\mathcal{X} = \mathbb{R}$ and the label space will be $\mathcal{Y} = \{0, 1\}$.

The hypothesis class is the class of indicator functions on intervals in $\mathbb{R}$

$$\mathcal{H} = \{\delta_I \,:\, I = [a, b]\}\,.$$

The $\mathbf{ERM}_{\mathcal{H}}$ rule will be implemented by finding the smallest interval containing all data points labeled 1. Above, $\mathbf{ERM}_{\mathcal{H}}(\mathcal{T}) = h_{\mathcal{T}}$.

Mango Space



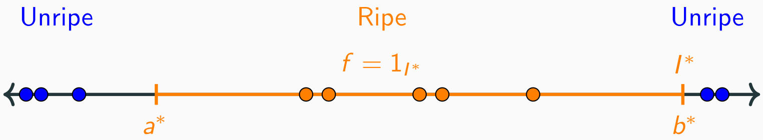
$$\mathcal{X} = \mathbb{R}$$

We will make two assumptions:

The I.I.D Assumption: The examples in the training set $\mathcal{T}$ are independent and identically distributed (i.i.d.). That is, the probability the i'th point in the training set taking on a certain value depends only on $\mathcal{D}$.

The Realizability Assumption: The labels are actually generated by some $f \in \mathcal{H}$ and drawn from some distribution $\mathcal{D}$ on $\mathcal{X}$.

Mango Space



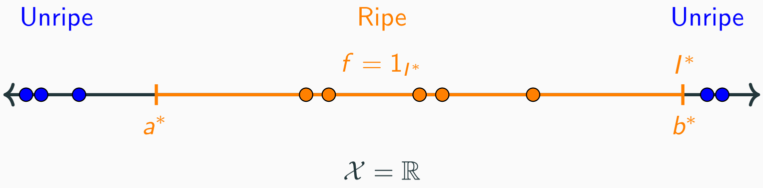
$\mathcal{X} = \mathbb{R}$

We will make two assumptions:

The I.I.D Assumption: The examples in the training set $\mathcal{T}$ are independent and identically distributed (i.i.d.). That is, the probability the i'th point in the training set taking on a certain value depends only on $\mathcal{D}$.

The Realizability Assumption: The labels are actually generated by some $f \in \mathcal{H}$ and drawn from some distribution $\mathcal{D}$ on $\mathcal{X}$.

Mango Space

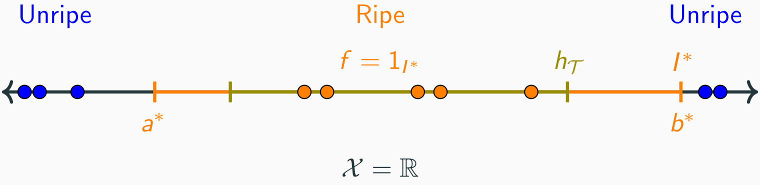


The Realizability Assumption: The labels are actually generated by some $f \in \mathcal{H}$ and drawn from some distribution $\mathcal{D}$ on $\mathcal{X}$.

Realizability implies that for any training set $\mathcal{T}$, any ERM rule will return a classifier $h_{\mathcal{T}}$ with **training error** $L_{\mathcal{T}}(h_{\mathcal{T}}) = 0$.

But that does not meant that ERM returns f! We are not interested in **training error**, we are interested in the **true error** $L_{(f,\mathcal{D})}(h_{\mathcal{T}})$.

Mango Space

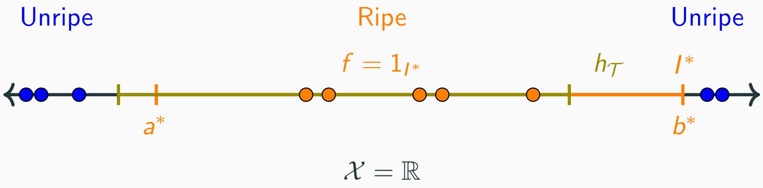


The Realizability Assumption: The data is actually generated by some $f \in \mathcal{H}$ and some distribution $\mathcal{D}$ on $\mathcal{X}$.

Realizability implies that for any training set $\mathcal{T}$, any ERM rule will return a classifier $h_{\mathcal{T}}$ with **training error** $L_{\mathcal{T}}(h_{\mathcal{T}}) = 0$.

But that does not meant that ERM returns f! We are not interested in **training error**, we are interested in the **true error** $L_{(f,\mathcal{D})}(h_{\mathcal{T}})$.

Mango Space

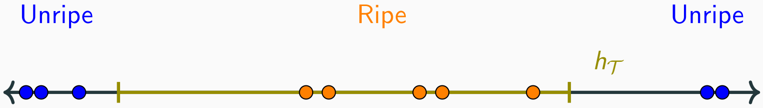


The Realizability Assumption: The data is actually generated by some $f \in \mathcal{H}$ and some distribution $\mathcal{D}$ on $\mathcal{X}$.

Realizability implies that for any training set $\mathcal{T}$, any ERM rule will return a classifier $h_{\mathcal{T}}$ with **training error** $L_{\mathcal{T}}(h_{\mathcal{T}}) = 0$.

But that does not meant that ERM returns f! We are not interested in **training error**, we are interested in the **true error** $L_{(f,\mathcal{D})}(h_{\mathcal{T}})$.

Mango Space



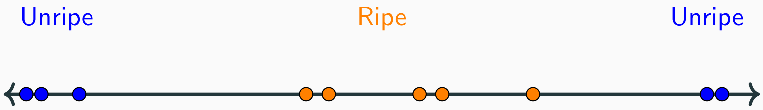
$$\mathcal{X} = \mathbb{R}$$

The Realizability Assumption: The data is actually generated by some $f \in \mathcal{H}$ and some distribution $\mathcal{D}$ on $\mathcal{X}$.

Realizability implies that for any training set $\mathcal{T}$, any ERM rule will return a classifier $h_{\mathcal{T}}$ with **training error** $L_{\mathcal{T}}(h_{\mathcal{T}}) = 0$.

But that does not meant that ERM returns f! We are not interested in **training error**, we are interested in the **true error** $L_{(f,\mathcal{D})}(h_{\mathcal{T}})$.

Mango Space



$$\mathcal{X} = \mathbb{R}$$

Deriving the Minimum Training Set Size:

Let $\epsilon \in (0,1)$ be the **target accuracy**, that is we interpret the event $L_{(\mathcal{D},f)}(h_{\mathcal{T}}) > \epsilon$ as a **failure** of the learning algorithm.

We want to derive an upper bound for the probability δ of picking a training set $\mathcal{T}$ with N elements such that $L_{(\mathcal{D},f)}(h_{\mathcal{T}}) > \epsilon$. We call $(1-\delta)$ the **confidence parameter**.

The goal is to find an upper bound for the probability

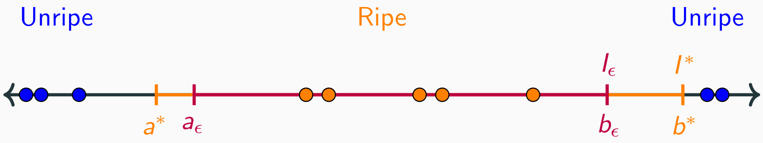
$$\delta = \mathcal{D}^N\big(\{\mathcal{T} \,:\, L_{(\mathcal{D},f)}(h_{\mathcal{T}}) > \epsilon\,\}\big)\,.$$

We will proceed in two steps:

First, we find a condition M on $\mathcal{T}$ that guarantees $L_{(\mathcal{D},f)}(h_{\mathcal{T}}) < \epsilon$. Preferably a condition that takes the form of a set of sets of training data, so that if $L_{(\mathcal{D},f)}(h_{\mathcal{T}}) < \epsilon$, then $\mathcal{T} \in M$.

Then, we derive an upper bound on the probability that $\mathcal{T} \notin M$.

Mango Space



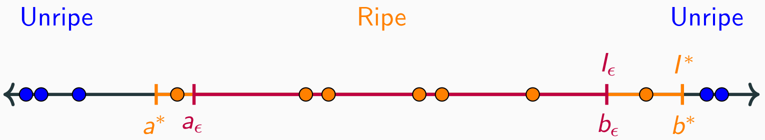
$$\mathcal{X} = \mathbb{R}$$

Let $I^* = [a^*, b^*]$ be the actual interval of ripeness. Pick $I_\epsilon = [a_\epsilon, b_\epsilon] \subset I^*$ so that both $\mathcal{D}([a^*, a_\epsilon]) = \epsilon/2$ and $\mathcal{D}([b_\epsilon, b^*]) = \epsilon/2$.

This means that the region $I^* - I_\epsilon$ contains less than ϵ of the total probability, ie $\mathcal{D}(I^* - I_\epsilon) = \epsilon$.

Then, if there exists $x_i, x_j \in \mathcal{T}$ such that $x_i \in [a_*, a_\epsilon]$ and $x_j \in [b_\epsilon, b_*]$, the ERM rule returns a classifier with error at most ϵ.

Mango Space



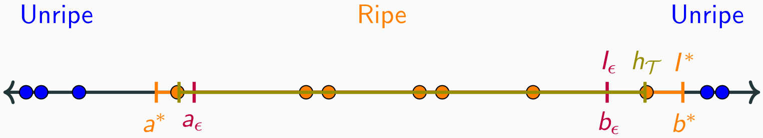
$\mathcal{X} = \mathbb{R}$

Let $I^* = [a^*, b^*]$ be the actual interval of ripeness. Pick $I_\epsilon = [a_\epsilon, b_\epsilon] \subset I^*$ so that both $\mathcal{D}([a^*, a_\epsilon]) = \epsilon/2$ and $\mathcal{D}([b_\epsilon, b^*]) = \epsilon/2$.

This means that the region $I^* - I_\epsilon$ contains less than ϵ of the total probability, ie $\mathcal{D}(I^* - I_\epsilon) = \epsilon$.

Then, if there exists $x_i, x_j \in \mathcal{T}$ such that $x_i \in [a_*, a_\epsilon]$ and $x_j \in [b_\epsilon, b_*]$, the ERM rule returns a classifier with error at most ϵ.

Mango Space



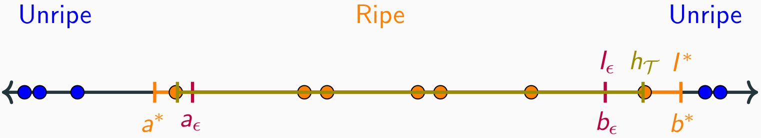
$\mathcal{X} = \mathbb{R}$

Let $I^* = [a^*, b^*]$ be the actual interval of ripeness. Pick $I_\epsilon = [a_\epsilon, b_\epsilon] \subset I^*$ so that both $\mathcal{D}([a^*, a_\epsilon]) = \epsilon/2$ and $\mathcal{D}([b_\epsilon, b^*]) = \epsilon/2$.

This means that the region $I^* - I_\epsilon$ contains less than ϵ of the total probability, ie $\mathcal{D}(I^* - I_\epsilon) = \epsilon$.

Then, if there exists $x_i, x_j \in \mathcal{T}$ such that $x_i \in [a_*, a_\epsilon]$ and $x_j \in [b_\epsilon, b_*]$, the ERM rule returns a classifier with error at most ϵ.

Mango Space



$$\mathcal{X} = \mathbb{R}$$

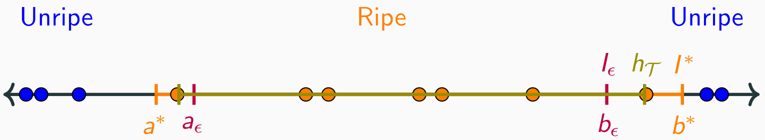
Formally,

$$\{x \,:\, h_{\mathcal{T}}(x) \neq f(x)\} \subseteq [a^*, a_\epsilon] \cup [b_\epsilon, b^*] \,,$$

so

$$\begin{aligned} L_{(\mathcal{D},f)}(h_{\mathcal{T}}) &= \mathcal{D}\big(\, \{x \,:\, h_{\mathcal{T}}(x) \neq f(x)\} \,\big) \\ &\leq \int_{a^*}^{a_\epsilon} \mathcal{D}(x)\, dx + \int_{b_\epsilon}^{b^*} \mathcal{D}(x)\, dx \\ &= \epsilon \,. \end{aligned}$$

Mango Space



$$\mathcal{X} = \mathbb{R}$$

We have shown that if a randomly selected training set $\mathcal{T}$ of size N contains a point in $[a^*, a_\epsilon]$ and a point in $[b_\epsilon, b^*]$ than the ERM classifier must have total error less than ϵ. What is the probability of this result **not** attaining?.

The probability of choosing N training point i.i.d. from $\mathcal{D}$ and having either none in $[a^*, a_\epsilon]$ or in $[b_\epsilon, b^*]$ is

$$\mathcal{D}^N\left(\left\{\mathcal{T} \,:\, x_i \notin [a^*, a_\epsilon], \forall\, i\right\} \cup \left\{\mathcal{T} \,:\, x_i \notin [b_\epsilon, b^*], \forall\, i\right\}\right) \geq \delta\,.$$

By the union bound,

$$\begin{aligned}
\mathcal{D}^N&\Big(\big\{\mathcal{T} : x_i \notin [a^*, a_\epsilon], \forall\, i\big\} \cup \big\{\mathcal{T} : x_i \notin [b_\epsilon, b^*], \forall\, i\big\}\Big) \\
&\le \mathcal{D}^N\Big(\big\{\mathcal{T} : x_i \notin [a^*, a_\epsilon], \forall\, i\big\}\Big) + \mathcal{D}^N\Big(\big\{\mathcal{T} : x_i \notin [b_\epsilon, b^*], \forall\, i\big\}\Big) \\
&= \prod_{i=1}^{N} \mathcal{D}\Big(\big\{x : x \notin [a^*, a_\epsilon]\big\}\Big) + \prod_{i=1}^{N} \mathcal{D}^N\Big(\big\{x : x \notin [b_\epsilon, b^*]\big\}\Big)
\end{aligned}$$

Where we have used the i.i.d. assumption to notice that joint distribution is the same as the product of N distributions.

Since

$$\mathcal{D}\Big(\{x : x \notin [a^*, a_\epsilon]\}\Big) = 1 - \frac{\epsilon}{2}\,,$$

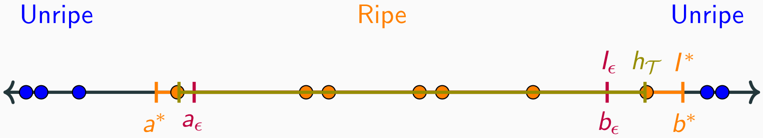
we can bound the probability by

$$\begin{aligned}
&\prod_{i=1}^{N}\mathcal{D}\Big(\{x : x \notin [a^*, a_\epsilon]\}\Big) + \prod_{i=1}^{N}\mathcal{D}^N\Big(\{x : x_i \notin [b_\epsilon, b^*]\}\Big)\\
&\qquad = (1-\epsilon/2)^N + (1-\epsilon/2)^N\\
&\qquad \leq 2e^{-N\cdot\epsilon/2}
\end{aligned}$$

The last is by Taylors approximation and is for connivance only. We have shown that

$$\delta \leq 2e^{-N\cdot\epsilon/2}\,.$$

Mango Space



$$\mathcal{X} = \mathbb{R}$$

A bit of algebra shows the following: For any $\delta > 0$ and any $\epsilon > 0$, assume we have

$$N \geq \frac{2\log(2/\delta)}{\epsilon}$$

points of the training data in $\mathcal{T}$. Then, with probability greater than $1-\delta$, the ERM rule returns a classifier $h_\mathcal{T}$ with total loss less than ϵ.
We say then that with enough data, the output rule is **probability approximately correct**.

References

3d pictures were made in GeoGebra (https://www.geogebra.org).

References for the probability review can be found at

http://cs229.stanford.edu/section/cs229-prob.pdf

http://www.cs.princeton.edu/courses/archive/spring07/cos424/scribe_notes/0208.pdf

This lecture covers chapter 2 of Shalevl-Shwarts and Ben-David.