UniTs - University of Trieste

Faculty of Scientific and Data Intensive Computing

Department of mathematics informatics and geosciences

Probabilistic Machine Learning

Lecturer:
Prof. Luca Bortolussi

Authors:
Andrea Spinelli
Christian Faccio

March 10, 2025

github.com/Spina02 andreaspinelli2002@gmail.com

Abstract

As a student of Scientific and Data Intensive Computing, I've created these notes while attending the **Probabilistic Machine Learning** course.

This course will cover the "*probabilistic side*" of machine learning. In particular, we will focus on the following topics:

- Basics of probability and probabilistic inference
- Pronabilistic formulation of learning (Empirical Risk Minimization and PAC Learning)
- Graphical Models
- Inference with graphical models: belief propagation
- Hidden Markov Models for sequential data
- Bayesian Linear Regression and Classification, Laplace approximation, Model Selection
- Kernel Regression and Kernel functions, Gaussian Processes for regression (hints)
- Monte Carlo sampling
- Expectation Maximization and Variational Inference
- Bayesian Neural Networks
- Generative Modelling: Variational Autoencoders and Diffusion Processes

While these notes were primarily created for my personal study, they may serve as a valuable resource for fellow students and professionals interested in probabilistic machine learning.

Contents

1 Introduction

1.1 Machine Learning

Machine learning is a field of computer science about **learning models**.

Models

Definition: ***Model***

- A ***Model*** is a hypothesis that certain features of a system of interest are well replicated in another, simpler syetem.
- ***Mathematical Model*** is a model where the simpler system consists of a set of mathematical relations between objects (equations, inequalities, etc).
- A ***Stochastic Model*** is a mathematical model where the objects are probability distributions.

All modelling usually starts by defining a family of models indexed by some parameters, which are tweaked to reflect how well the feature of interest are replicated.
Machine learning deals with algorithms for automatic selecrion of a model from observations of the system.

Machine Learning

Definition: ***Machine learning***

Machine learning *explores the study and construction of algorithms that can learn from and make predictions on data.* *Wikipedia*

There are three main types of machine learning:

Generative and Discriminative Learning

- **Generative Learning** im at describing the full probability distribution of inputs x or input/output pairs (x,y).
$$p(x,y) = p(x)p(y|x)$$
- **Discriminative Learning** aims at describing the conditional probability of output given the input, or a statistics/function of such probability
$$p(y|x) \quad or \quad y = f(x)$$

[to fix:]
- **Supervised Learning**: The algorithm learns from labeled data by mapping inputs to outputs.
- **Unsupervised Learning**: The algorithm identifies patterns or structures in unlabeled data.
- **Data Generatiion**: The algorithm generates new data points.

Inference and Estimation

Two central concepts for probabilistic machine learning are:

- **Inference**: Compute marginals and contitionals probability distributions applying the laws of probability.
- **Estimation**: Given data and a family of models, find the best parameters/models that explains the data.

In the Bayesian world: estimation $\approx$ inference.

Probability

Probability is a mathematical theory that deals with **uncertainty**
When a certain problems has to face practical difficulties due to it's complexity, we can use probability to model the ***aleatorical uncertainty***, which is the uncertainty due to the randomness of the system.
More often, we have a limited knowledge of the system, and we can use probability to model the ***epistemic uncertainty***, which is the uncertainty due to the lack of knowledge.

Tip: ***Everything is a probability distribution***

In machine learning **everything is a probability distribution**, even if not explicitly stated.

1.2 Probability basics

1.2.1 Random Variables

Random Variables are functions mapping outcomes of an experiment to real numbers. They serve as abstract representations of the outcomes in randomized experiments. Note that what we observe are the *realizations* (values resulting from an observed outcome) of these random variables.

Example: ***Random Variable***

Consider the following example:

$$\{Head, Tail\}, \qquad \{0,1\}, \qquad \left\{\tfrac{1}{2}, \tfrac{1}{2}\right\}.$$

Only the second is the random variable itself; the third is its probability distribution, while the first is the sample space of potential outcomes.

We consider a ***Sample Space*** Ω, which is the set of all possible outcomes of a random experiment. A random variable X is a function:

$$X : \Omega \to E, \qquad \text{where } E \subseteq \mathbb{R} \quad (\text{or } E \subseteq \mathbb{N})$$

with the probability measure

$$P(X \in S) = P(\{\omega \in \Omega \mid X(\omega) \in S\}), \quad S \subseteq E.$$

A model for our random outcome is the probability distribution of X. In particular, if the sample space is finite or countable the **probability mass function (pmf)** is given by:

$$p(x) := P(X = x).$$

If the sample space is infinite, we use the **probability density function (pdf)** where

$$P(a \leq X \leq b) = \int_a^b p(x)dx \quad \text{and} \quad \int_{\mathbb{R}} p(x)dx = 1.$$

1.2.2 Notable Probability Distributions

Below are some of the most common probability distributions.

Discrete Distributions

Distribution	pmf	Mean	Variance
Binomial Bin(n,p)	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$
Bernoulli Bern(p)	$p \quad (x=1), \quad 1-p \quad (x=0)$	p	$p(1-p)$
Discrete Uniform $\mathscr{U}(a,b)$	$\dfrac{1}{b-a+1}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a+1)^2-1}{12}$
Geometric Geom(p)	$(1-p)^{x-1}p$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
Poisson Pois(λ)	$\dfrac{\lambda^x e^{-\lambda}}{x!}$	λ	λ

Continuous Distributions

Distribution	pdf	Mean	Variance
Continuous Uniform $\mathscr{U}(a,b)$	$\begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Exponential Exp(λ)	$\lambda e^{-\lambda x}$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$
Gaussian $\mathscr{N}(\mu,\sigma^2)$	$\dfrac{1}{\sigma\sqrt{2\pi}}\exp\left(-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right)$	μ	σ^2
Beta Beta(α,β)	$\dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	$\dfrac{\alpha}{\alpha+\beta}$	$\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Gamma Gamma(α,β)	$\dfrac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$	$\dfrac{\alpha}{\beta}$	$\dfrac{\alpha}{\beta^2}$
Dirichlet $Dir(\alpha)$	$\dfrac{1}{B(\alpha)}\prod_{i=1}^{K}x_i^{\alpha_i-1}$	$\tilde{\alpha}_i$	$\dfrac{\tilde{\alpha}_i(1-\tilde{\alpha}_i)}{\alpha_0+1}$
Student's t $St(\nu)$	$\dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\dfrac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	0	$\begin{cases} \frac{\nu}{\nu-2} & \nu > 2 \\ \infty & 1 < \nu \le 2 \end{cases}$

Notes:

- For discrete distributions, $n \in \{0,1,2,\dots\}$, $p \in [0,1]$, and x runs over the support.
- For continuous distributions, parameters such as λ, μ, σ, α, and β belong to $\mathbb{R}$ (with appropriate restrictions) and $x \in \mathbb{R}$.
- In the Dirichlet distribution, $\tilde{\alpha}_i = \frac{\alpha_i}{\sum_{h=1}^{K}\alpha_h}$ and $\alpha_0 = \sum_{i=1}^{K}\alpha_i$.

- For Student's t-distribution, $\nu > 1$.

2

Empirical Risk Minimization and PAC Learning

In this chapter, we will introduce the concept of **Empirical Risk Minimization** (ERM) in which to frame learning problems, the notion of inductive bias, and the main results of algorithmic learnability, encapsulated in the definition of **Probably Approximately Correct** (PAC) Learning and of complexity of a set of hypothesis, namely VC-dimension and Rademacher complexity.

2.1 Empirical Risk Minimization

We begin by considering a supervised learning setting in which the **input space** X is a subset of $\mathbb{R}^n$, and the **output space** Y can be real-valued (e.g., $Y = \mathbb{R}$), binary (e.g., $Y = \{0,1\}$), or a finite set of classes (e.g., $Y = \{0,1,\ldots,K\}$). In this probabilistic framework, each input-output pair (x,y) is drawn from a joint probability distribution

$$p(x,y) \in \text{Dist}(X \times Y),$$

often referred to as the ***data generating distribution***.
By definition, this distribution factors into the marginal $p(x)$ and the conditional $p(y \mid x)$, so that

$$p(x,y) = p(x)\,p(y \mid x).$$

Because $p(x)$ and $p(y \mid x)$ describe how inputs and outputs are related, it is helpful to write them explicitly. The marginal distribution of x is

$$p(x) = \int p(x,y)\,dy,$$

while the conditional distribution of y given x is

$$p(y \mid x) = \frac{p(x,y)}{p(x)}.$$

A typical dataset D in supervised learning consists of N input-output pairs drawn independently from $p(x,y)$. We denote this as

$$D \sim p^N(x,y),$$

which means

$$D = \{(x_i,y_i) \mid i = 1,\ldots,N\},$$

where each (x_i,y_i) is sampled according to the joint distribution $p(x,y)$.
In many cases, we assume that $p(y \mid x)$ depends on some unknown function of x. Formally, one might write

$$p(y \mid x) = p\big(y \mid f(x)\big),$$

where f is the function we aim to learn. The central objective in supervised learning—through methods such as empirical risk minimization—is to find or approximate this function f by using the observed data D.

2.2 Risk and Empirical Risk

$h \in \mathscr{H} \quad x,y \sim p(x,y)$
loss function $l(x,y,h) \in \mathbb{R}_{\geq}$,
- 0-1 loss: $l(x,y,h) = \mathbb{I}(h(x) \neq y)$, eith $y \in \{0,1\}$.
- squared loss: $l(x,y,h) = (h(x)-y)^2$, with $y \in \mathbb{R}$.

We have a probabilistic process, so we have some inputs that are more likely than others. If a model makes a mistake on a more likely input, it should be penalized more.

Definition: ***Risk***

The ***risk*** (or ***generalization error***) is defined as:

$$R(h) = E_{x,y\sim p(x,y)}[l(x,y,h)]$$

Risk minimization principle:
The goal is to find the hypothesis h that minimizes the risk.

$$\text{find } h^* \in \mathscr{H} \text{ such that } h^* = \arg\min_{h\in\mathscr{H}} R(h)$$

Definition: ***Empirical Risk***

The ***empirical risk*** (or ***training error***) is defined as:

$$\hat{R} = \frac{1}{N}\sum_{i=1}^{N} l(x_i,y_i,h)$$

Empirical risk minimization principle:
The goal is to find the hypothesis h that minimizes the empirical risk.

$$\text{find } h_D^* = \arg\min_{h\in\mathscr{H}} \hat{R}(h)$$

2.2.1 Bias Variance Trade-off

In this section, we want to analyze the generalization error and decom- pose it according to the sources of error that we are going to commit.
In what follows, we will use the squared loss (hence we will focus on regression problems). Considering $h \in \mathscr{H}$, an explicit expression of the generalization error committed when choosing hypothesis h is:

$$R(h) = E_p[l(x,y,h)] = \int\int (h(x)-y)^2 p(x,y)dxdy$$

Theorem 1. *The minimizer of the generalization error R is:*

$$\boxed{g(x) = E[y|x] = \int y p(y|x)dy}$$

so that $g = \arg\min_h R(h)$, *if* $g \in \mathscr{H}$
We can rewrite the risk as:

$$R(h) = \underbrace{\int (h(x)-g(x))^2 p(x)dx}_{=0 \quad iff \quad h(x)=g(x)} + \overbrace{\iint (g(x)-y)^2 p(x,y)dxdy}^{\text{independent of } h\text{: intrinsic noise}}$$

$$
\begin{aligned}
E_D[R(h_D^*)] &= \underbrace{\int (E_D[h_D^*(x) - g(x)])^2 p(x)dx}_{bias^2} \\
&+ \underbrace{\int E_D[(h_D^*(x) - E_D[h_D^*(x)])^2 p(x)dx]}_{variance} \\
&+ \underbrace{\iint (g(x) - y)^2 p(x,y)dxdy}_{noise}
\end{aligned}
$$

2.3 ERM and Maximum Likelihood

Given a dataset $D = \{(x_i, y_i)\}_{i=1,\dots,m}$ s.t. $D \sim p^m, p = p(x,y)$
We factorize the data generating distributions as: $p(x,y) = p(x)p(y|x)$ and we make an hypothesis on $p(y|x)$, trying to express this conditional probability in a parametric form:

$$p(y|x) = p(y|x,\theta)$$

[check recording for missing part]
We consider the log Likelihood:

$$L(\theta; D) = \sum_{i=1}^{m} \log p(y_i|x_i,\theta)$$

Then we apply the maximum likelihood principle:

$$
\begin{aligned}
\arg\min_\theta - L(\theta; D) &= \arg\min_\theta - \frac{1}{m} \sum_{i=1}^{m} \log p(y_i|x_i,\theta) \\
&= \arg\min_\theta E_{p(x,y)}[-\log p(y|x,\theta)]
\end{aligned}
$$

since the avarage is an empirical approximation of the expectation.

Observation: ***Empirical Risk***

$-\frac{1}{m} \sum_{i=1}^{m} \log p(y_i|x_i,\theta)$ is known as **empirical risk**

2.4 KL divergence

Consider a probability distribution $p(x)$, then $-\log p(x)$ is a measure of **self-information**. Indeed, if $p(x) = 1$ then $-\log p(x) = 0$ (no self-information), describing substantially out (lack of) surprise in observing the event. If instead $p(x) = 0$ then $-\log p(x) = \infty$. In general, the more rare the event is, i.e. the lower is $p(x)$, the more self-information it carries, i.e. the larger is $-\log p(x)$.

Definition: ***Entropy***

In an information-theoretic sense, the **entropy** is a measure of the inforamtion that is carries by a random phenomenon, expressed as the expected amount of self-information that is conveyed by a realization of the random phenomenon.

It is formally defined as:

$$H(p) = E_p[-\log p(x)] = -\int p(x)\log p(x)dx$$

for the continuous case, and:

$$H(p) = E_p[-\log p(x)] = -\sum p(x)\log p(x)$$

for the discrete case.

For the discrete case, the maximum entropy is achieved for te uniform distrivution and is equal to $\log n$, where n is the number of possible outcomes. In the continuous case, for a fixed variance, the distribution that maximizes the entropy is the Gaussian. The entropy is always 0 if we have a deterministic distribution.

Definition: ***Kullback-Leibler divergence***

The **Kullback-Leibler divergence** (or **relative entropy**) between two probability distributions $p(x)$ and $q(x)$ is a measure of how one distribution diverges from a second, expected probability distribution.
It is formally defined as:

$$D_{KL}(p||q) = E_p\left[\log\frac{p(x)}{q(x)}\right] = \int p(x)\log\frac{p(x)}{q(x)}dx$$

for the continuous case, and:

$$D_{KL}(p||q) = E_p\left[\log\frac{p(x)}{q(x)}\right] = \sum p(x)\log\frac{p(x)}{q(x)}$$

for the discrete case.

Intuitively, we are taking a sort of expected difference between p and q, expressed in terms of a log odds ratio. It tells us how different two distributions are: the larger KL the more different are p and q.
Properties of KL:

- $KL[q||p]$ is a convex function of q and p and $KL[q||p] \geq 0$
- KL is non-symmetric, i.e. $KL[q||p] \neq KL[p||q]$
- $KL[q||p] = -H[q] - \mathbb{E}_q[\log p]$, where the first term is the entropy and the second term is known as cross-entropy between p and q.

Suppose moreover, that p is fixed but unknown, $q = q_0$ can vary: what we usually do is trying to find the best q_θ that approximates p.
The **mutual information** between x and y is defined as:

$$I(x,y) = KL[p(x,y)||p(x)p(y)] = \int\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy$$

$KL[p(x,y)||p(x)p(y)] = 0$ iff x and y are independent.
Moreover, the more dependent they are, the more different is $p(x,y)$ from the product of the marginals, the more information x carries about y and viceversa.
In other words, the higher the mutual information is, the more knowing y will tell us about x, the less residual uncertainty on x we will have.

Consider a dataset: $\underline{x} : x_1, \ldots, x_N$:

Definition: *Empirical distribution*

The **empirical distribution** of a dataset $\underline{x}$ is defined as:

$$\hat{p}(\underline{x}) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i)$$

where $\delta(x)$ is the Dirac delta function.

It is an approximation of the input data generating function$p(x)$. Practically, the more observations we have, the more the empirical distribution will look like $p(x)$.
Given a distribution q, we can compute:

$$KL[p_{emp}||q] = \mathbb{E}_{p_{emp}}\left[\log \frac{p_{emp}(x)}{q(x)}\right] = \frac{1}{N}\sum_{i=1}^{N} \log \frac{p(x_i)}{q(x_i)}$$

If $q = q_0$, this is $-\frac{1}{N}L(\Theta)$ plus a constant. Hence maximizing $L(\Theta)$ is essentially equivalent to minimizing the KL between p_{emp} and q_0. This means that we can always rephrase maximum likelihood in terms of cross-entropy.

2.5 PAC Learning

Our goal is to measure how much we can learn as a function of the model complexity. This results in the **PAC** (Probably Approximately Correct) **learning** framework, which encodes the notion of model complexity and gives also bounds on the error that we commit. Here we consider it in the context of (binary) classification, i.e. $y \in 0, 1$, using the 0-1 loss.

Definition: *PAC Learning*

A realizable hypothesis set $\mathscr{H}$ is ***PAC learnable*** iff $\forall \varepsilon, \delta \in (0,1), \forall p(x,y), \exists m_{\varepsilon,\delta} \in \mathscr{N}\ s.t. \forall m \geq m_{\varepsilon,\delta}, \exists D\ p^m, |D| = m$ then $p_D(R(h_D^*) \leq \varepsilon) \geq 1 - \delta$

This means that, fixing two parameters $\varepsilon, \delta \in (0,1)$, governing our precision, and a data generating distribution $p(x,y)$, we can find a number of samples $m_{\varepsilon,\delta}$ such that, with probability at least $1-\delta$, the empirical risk of the best hypothesis h_D^* is less than ε. Note that the probability here is over the dataset D, meaning that our learning will succeed for a fraction $1-\delta$ of sampled datasets.
In a more general setting,

Definition:

Given an hypothesis set $\mathscr{H}$ (not necessarily realizable) and an algorithm A, $\mathscr{H}$ **is agnostic PAC-learnable** $if f \forall \varepsilon, \delta \in (0,1), \forall p(x,y), \exists m_{\varepsilon,\delta} \in \mathscr{N}\ p^m, |D| = m \geq m_{\varepsilon,\delta} \Rightarrow p_D(R(h_D^*) \leq R(h^*) + \varepsilon) \geq 1 - \delta$, being h_D^A the result of applying A to $\mathscr{H}$ and D.

This means that, fixing two parameters $\varepsilon, \delta \in (0,1)$, governing our precision, and a data generating distribution $p(x,y)$, we can find a number of samples $m_{\varepsilon,\delta}$ such that, with probability at least $1-\delta$, the empirical risk of the best hypothesis h_D^* is less than the risk of the best hypothesis h^* plus ε.

Finite hypothesis sets

An hypothesis set is said to be **finite** is $\mathscr{H}$ is s.t. $|\mathscr{H}| < \infty$.

Using combinatorial arguments, we can prove that finite hypothesis sets are agnostic PAC-learnable with:

$$m_{\varepsilon,\delta} \leq \lceil \frac{2\log(\frac{2\mathscr{H}}{\delta})}{\varepsilon^2} \rceil$$

hence with polynomial dependency on ε and δ. In this framework, $\log(|\mathscr{H}|)$ is a measure of the complexity of the set $\mathscr{H}$.

Warning:

If $\mathscr{H}$ is described by d parameters of type double when represented in a computer (64 bits), it holds that $|\mathscr{H}| \leq 2^{d64}$, so we have a finite set of hypothesis, hence we can provide a bound on every implementable set of hypothesis functions.
In this case,

$$m_{\varepsilon,\delta} \leq \frac{128d + 2\lg(\frac{2}{\delta})}{\varepsilon^2}$$

i.e. we have linear dependency on the number of parameters.

TO FINISH PAC LEARNING EXAMPLE...

2.6 VC Dimension (Vapnik-Chervonenkis)

3
Probabilistic Graphical Models