MATH 412: STATISTICAL MACHINE LEARNING

Rafiki's Notes

Rafael Barroso

Ingenierie Mathematique

École Polytechnique Fédérale de Lausanne

September 16, 2025

Contents

1 Supervised Learning Basics

In this area of machine learning, we try to understand certain relations beween input-output data. If such relations are established, we then wish to generalize for new *unseen* data. Things start getting even jucier whenever we wish to take decisions based on new data, resulting in a more generalized task. In this section we explore the formalization provided by Prof. Zemel.

1. **We have:**

   - Data: $\mathcal{D}_n := \{(x_0, y_0), \ldots, (x_n, y_n)\}$
   - i.e. tuples of the form (x_i, y_i)
   - $x_i :=$ input; $y_i :=$ output

2. **We want:**

   - Given $\mathcal{D}_n$, learn relations of the x_i's with the corresponding y_i's such that we may infer something about a new unseen y' given x'.

We now define the two types of tasks cosidered inside supervised learning (amongst others).

> **Definition 1.1.** A *prediction* task is established to be the discovery of y' (unseen) given x'. A *decision* task on the other hand, focuses on producing a decision based on (x', y') only with the data of x'

For example, take into consideration a medical diagnosis. We have $x_i :=$ patient data i.g. $\{\text{weight}_i, \text{height}_i, \ldots\}$; $y_i := \{\text{positive}, \text{negative}\}$. Then, a **prediction task** would consist in predicting y' given x'. A **decision task** on the other hand, would then consist on choosing how to treat patient x' i.g. choosing medicine $m \in \{A, B, C\}$ (we have to decide on y' by only seeing x').

We now consider the space of all possible decisions; a *learning algorithm* (sometimes called *learning scheme*) $\mathscr{A}$.

> **Definition 1.2.** We define a learning algorithm as
> $$\mathscr{A} : \mathcal{D}_n \to \hat{f}$$
> where $\hat{f}$ is our decision function.

Obviously we want $\hat{f}$ to be "good" (otherwise, *nos estamos haciendo pendejos*). Hence, we must define what it means for $\hat{f}$ to be "good" i.e. what we want from $\hat{f}$.

Definition 1.3. Let $\mathcal{X}$ be the input space, then, a decision function is defined as

$$f : \mathcal{X} \to \mathcal{A}^{\mathcal{X}}$$

Note that the input space $\mathcal{X}$ is the space of all x_i's.

Ideally, as stated before, we want a "good" function (i.e. decision function) f such that $f(x) \in \mathcal{A}^{\mathcal{X}}$ is "good" when compared to an unseen y. This means that $f(x)$ must be an accurate prediction of y and it has the **smallest possible cost** whenever y occurs. So, we compute the *loss function l*.

Definition 1.4. Let $\mathcal{Y}$ be the space of all possible outcomes, then

$$l : \mathcal{A}^{\mathcal{X}} \times \mathcal{Y} \to \mathbb{R}$$

defined by $(f(x) = a, y) \mapsto l(a, y)$. Note that this function meassures the cost of taking decision f whenever y occurrs i.e. the *risk*.

Remark 1.5. Note that all the above definitions boil down to the fact that we're trying to design a 'good' learning algorithm $\mathscr{A}$ that produces $\hat{f}$ in such a way that the risk is minimized. We formalize the definition of a learning algorithm as follows

$$\mathscr{A} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{A}^{\mathcal{X}} \text{ given by } \mathcal{D}_n \mapsto \hat{f}$$

Throughout the lecture, unless stated otherwise, we assume that the data is generated by a stochastic process and done so i.i.d. as random variables i.e. (X_i, Y_i).

Because of the fact that this is a statistics class, we'll start getting into the *deets* using much more of their language (statisticians have a fetish for fancy language and syntax). Hence we would like to define what the *expected cost* of taking decision f as the risk $\mathcal{R}$.

Definition 1.6. We define the risk as follows

$$\mathcal{R}(f) := \mathbb{E}\big[l(a, Y)\big]. \text{ If } \exists f^* \in \mathcal{A}^{\mathcal{X}} : \mathcal{R}(f^*) = \inf_{f \in \mathcal{A}^{\mathcal{X}}} \mathcal{R}(f)$$

then, that f^* is our juicy function we're looking for! statisticians call it the *target* function. Now, the *conditional risk* of taking f as an action given x has happened is defined as

$$\mathcal{R}(f(x) = a|x) = \mathbb{E}\big[l(a, Y)|X = x\big] = \int l(a, y) dP_{Y|X}(y|x)$$

Note that $dP_{Y|X}(y|x)$ just means we're integrating over the conditional distribution of Y given $X = x$. To simplify further (and remark the fetish statisticians posse), this just means that we're taking average the loss over all possible outcomes of Y, weighted by how likely they are given $X = x$.

Remark 1.7. Note that

$$\mathbb{E}\big[\mathcal{R}(f(X)|X)\big] = \mathbb{E}\big[\mathbb{E}\big[l(f(X),Y)|X=x\big]\big] = \mathbb{E}\big[l(f(X),Y)\big]$$

i.e. the expected value of $\mathcal{R}(f)$.

We finally make the last definition of the section, since we're interested in meassuring risks we shall compute the *excess risk* $\varepsilon(f)$ while we're at it. This number tells us how much of our risk is over the optimal ammount.

Definition 1.8. The excess risk is given by

$$\varepsilon(f) := \mathcal{R}(f) - \mathcal{R}(f^*) = \mathbb{E}\big[l(f(X),Y)\big] - \mathbb{E}\big[l(f^*(X),Y)\big]$$

$$\Rightarrow \varepsilon(f) = \mathbb{E}\big[l(f(X),Y) - l(f^*(X),Y)\big]$$

1.1   Examples

As for every new concept, one must get their hands dirty in order to fully grasp the idea that is being transmitted. We explore some basic examples as to how these things might be used in practice by taking a look at two main cases: *Ordinary Least Squares* and *Classification*. Our main goal is to derived the target function f^* given a loss function l.

1.1.1   Ordinary Least Squirts Regression (OLS)

Consider the case where both our action space $\mathcal{A}^{\mathcal{X}}$ and our output space $\mathcal{Y}$ are both the real numbers $\mathbb{R}$. The loss function in this case is then defined as $l : \mathbb{R}^2 \to \mathbb{R}$ and it is given by $l(f(x) = a, y) = (a - y)^2$. Then, the expected cost of taking decision f is the following

$$\mathcal{R}(f) = \mathbb{E}\big[l(a,y)\big] = \mathbb{E}\big[(a-y)^2\big] = \mathbb{E}\big[(y-a)^2\big]$$

We now make an educated guess (cheeky little trick) and assume that $\hat{f} = \mathbb{E}\big[Y|X\big]$ and so, we consider the following expected conditional risk:

$$\mathcal{R}(f(X)|X) = \mathbb{E}\big[l(a,y)|X\big] = \mathbb{E}\big[(Y-f(X))^2|X\big] = \mathbb{E}\big[(Y-\hat{f}(X)+\hat{f}(X)-f(X))^2|X\big]$$

Note that we also implemented a cheeky trick, it's just adding a glorified zero into the mix i.e. $-\hat{f}(X) + f^*(X) = 0$. Expanding the quadratic term inside the expectation, we obtain the following equation:

$$\mathbb{E}\big[(Y-\hat{f}(X))^2 + 2\big(Y-\hat{f}(X)\big)\big(\hat{f}(X)-f(X)\big) + (\hat{f}(X)-f(X))^2|X\big]$$

$$= \mathbb{E}\big[(Y-\hat{f}(X))^2|X\big] + \mathbb{E}\big[2\big(Y-\hat{f}(X)\big)\big(\hat{f}(X)-f(X)\big)|X\big] + \mathbb{E}\big[(F^*(X)-F(X))^2|X\big]$$

$$= \mathbb{E}\big[(Y-\hat{f}(X))^2|X\big] + 2\big(\hat{f}(X)-f(X)\big)\mathbb{E}\big[Y-\hat{f}(X)|X\big] + \mathbb{E}\big[(\hat{f}(X)-f(X))^2|X\big]$$

Remark 1.9. Note that $2\big(\hat{f}(X) - f(X)\big)\mathbb{E}\big[Y - \hat{f}(X)|X\big] = 0$ since

$$\mathbb{E}\big[Y - \hat{f}(X)|X\big] = \mathbb{E}\big[Y - \mathbb{E}[Y|X]|X\big] = \mathbb{E}[Y|X] - \mathbb{E}[Y|X] = 0$$

So, we get that the conditional risk:

$$\mathcal{R}(f(X)|X) = \mathbb{E}\big[(Y - \hat{f}(X))^2|X\big] + \mathbb{E}\big[(\hat{f}(X) - f(X))^2|X\big]$$

$$\Rightarrow \mathcal{R}(f(X)|X) = \mathcal{R}(\hat{f}(X)|X) + \mathbb{E}\big[(\hat{f}(X) - f(X))^2|X\big]$$

Now, let's think for a moment. We essentially want to find f^* (target function) such that our risk $\mathcal{R}(f(X)|X)$ is minimized whenever we take that action.

Since $\mathcal{R}(\hat{f}(X)|X)$ can't help us, we turn our but cheeks to the other term, i.e. $\mathbb{E}\big[(\hat{f}(X) - f(X))^2|X\big]$. Note that if $f = \hat{f}$, this term goes to zero, and hence our conditional risk is minimized. Voila! since there actually exists such function that minimizes our risk, we've arrived at our sensual (target) function babyyy.

$$\therefore \; f^* = \hat{f}$$

1.1.2 Classification

We now consider a classification problem where $\mathcal{A}^{\mathcal{X}} = \mathcal{Y} := \{0, \ldots, K - 1\}$ and the loss function is given by the 0-1 loss function i.e. $l(a, y) = \mathbb{1}_{\{a \neq y\}}$. Then, the risk of f is given by the equation

$$\mathcal{R}(f) = \mathbb{E}\big[l(f(x) = a, y)\big] = \mathbb{P}\big(f(X) \neq Y\big) = 1 - \mathbb{P}\big(f(X) = Y\big)$$

Now, if we consider the conditional risk, we obtain the following derivation

$$\mathcal{R}(f|X = x) = \mathbb{E}\big[l(f(x) = a, y)|X = x\big] = \mathbb{P}\big(f(x) \neq Y|X = x\big) = 1 - \mathbb{P}\big(f(x) = Y|X = x\big)$$

Think for a moment. This states that in order to minimize $\mathcal{R}(f|X = x)$ we must then find the highest possible value for $\mathbb{P}\big(f(x) = Y|X = x\big) = \mathbb{P}\big(Y = f(x)|X = x\big)$. So, the following equivalence holds

$$\min \mathcal{R}(f|X = x) = \max_{f \in \mathcal{A}^{\mathcal{X}}} \mathbb{P}\big(Y = f(x)|X = x\big)$$

$$\Rightarrow \arg \max_{k \in K} \mathbb{P}\big(Y = k|X = x\big) = f^*$$

And so, we've found our hard-to-get (like women playing their games) target function. That last implication basically states that we must do the following: *for each value of x, pick y that makes it the largest possible value.*

This is a great book![1]

6

Exercise Sheet 1

The following exercises are designed to reinforce the concepts introduced in this section. They provide both practice with fundamental techniques and opportunities to explore some extensions of the main results. Complete solutions are included to aid understanding and self-study.

Exercise 1.1: Classification from a discrete input space

We consider a multiclass classification problem with 3 classes $\mathcal{Y} = \{1, 2, 3\}$ for data with only a single discrete descriptor in $\mathcal{X} = \{1, 2, 3, 4\}$.

We assume that the joint probability distribution $\mathbb{P}(Y = y, X = x)$, with $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, is specified by the following table:

	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1$	0.02	0.08	0.10
$X = 2$	0.05	0.40	0.15
$X = 3$	0.02	0.02	0.12
$X = 4$	0.02	0.01	0.01

1. What is the target function f^* for the 0–1 loss?

2. What are the values of $f^*(x)$ for $x = 1, 2, 3, 4$?

3. What is the value of the risk for the target function?

Answers:

As shown above, $f^*(x) = \arg\max_{y \in \mathcal{Y}} \mathbb{P}(Y = y | X = x)$.

Now, if $\mathcal{X} = \{1, 2, 3, 4\}$ and $\mathcal{Y} = \{1, 2, 3\}$, where $\mathbb{P}(Y = y | X = x)$ is given by a table, we find $f^* * (x)$ for $x \in \mathcal{X}$. We want to find the y that maximizes $\mathbb{P}(Y = y | X = x)$. For example, if $\max_y \mathbb{P}(Y = y | X = 1) = 0.1$, which occurs when $y = 3$, then $f^*(1) = 3$. So, for $x = 1, f^*(1) = 3$. If $x = 2, f^*(2) = 2$. For $x = 3, f^*(3) = 3$. For $x = 4, f^*(4) = 1$.

Finally, the risk is the probability of misclassification:

$$\mathbb{P}(f^*(X) \neq Y) = \mathbb{E}[\mathbb{1}_{f^*(X) \neq Y}] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathbb{1}_{f^*(x) \neq y} \mathbb{P}(X = x, Y = y)$$

Exercise 1.2: Recap of linear models

Let $y = X\beta + \varepsilon$, where $\mathbb{E}[\varepsilon] = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 I$, and X is a non-random full rank matrix of size $n \times p$. This setup contains the Gauss–Markov assumptions of a linear model.

1. Derive the least squares estimator $\hat{\beta} = (X^\top X)^{-1} X^\top y$.

2. Show that $\hat{\beta}$ is unbiased and that the variance of $\hat{\beta}$ is given by $\sigma^2 (X^\top X)^{-1}$.

Answers:

The model is $y = X\beta + \epsilon$, with assumptions $\mathbb{E}[\epsilon] = 0$ and $\text{Var}(\epsilon) = \sigma^2 I$. We assume X is full rank. Note that $X^T X$ is therefore invertible (positive definite). Since $\mathbb{E}[\epsilon] = 0$, we have $\mathbb{E}[Y|X] = X\beta$.

Now, the loss function is $L(\beta) = \sum_i (y_i - x_i^T \beta)^2 = (y - X\beta)^T (y - X\beta)$.

$$\begin{aligned}
L(\beta) &= (y^T - (X\beta)^T)(y - X\beta) \\
&= y^T y - y^T X\beta - (X\beta)^T y + (X\beta)^T (X\beta) \\
&= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \\
&= y^T y - 2\beta^T X^T y + \beta^T X^T X\beta
\end{aligned}$$

To find the minimum, we take the derivative with respect to β and set it to 0.

$$\frac{\partial L}{\partial \beta} = -2X^T y + 2X^T X\beta$$

Setting to 0:

$$-2X^T y + 2X^T X\beta = 0 \implies X^T X\beta = X^T y \implies \hat{\beta} = (X^T X)^{-1} X^T y$$

Now, let's find the expectation of the estimator $\hat{\beta}$ to check for bias.

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(X^T X)^{-1} X^T y] \\
&= (X^T X)^{-1} X^T \mathbb{E}[y] \\
&= (X^T X)^{-1} X^T \mathbb{E}[X\beta + \epsilon] \\
&= (X^T X)^{-1} X^T (X\beta + \mathbb{E}[\epsilon]) \\
&= (X^T X)^{-1} (X^T X)\beta + 0 \\
&= \beta
\end{aligned}$$

Therefore, $\mathbb{E}[\hat{\beta}] = \beta$, so the estimator is unbiased. Continuing from Ex 1.2, we find the variance of $\hat{\beta}$.

$$\begin{aligned}
\text{Var}(\hat{\beta}) &= \text{Var}((X^T X)^{-1} X^T y) \\
&= \text{Var}((X^T X)^{-1} X^T (X\beta + \epsilon)) \\
&= \text{Var}((X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T \epsilon) \\
&= \text{Var}(\beta + (X^T X)^{-1} X^T \epsilon) \\
&= 0 + \text{Var}((X^T X)^{-1} X^T \epsilon) \quad \text{(since } \beta \text{ is a constant)}
\end{aligned}$$

Using the rule $\text{Var}(AZ) = A\text{Var}(Z)A^T$:

$$
\begin{aligned}
\text{Var}(\hat{\beta}) &= ((X^TX)^{-1}X^T)\text{Var}(\epsilon)((X^TX)^{-1}X^T)^T \\
&= ((X^TX)^{-1}X^T)(\sigma^2 I)(X(X^TX)^{-1}) \\
&= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}I \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}
$$

Exercise 1.3: Linear regression for binary classification

Consider a binary classification problem with $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathcal{A} = \{-1, 1\}$. We model the conditional expectation of Y given X by the linear model

$$
\mathbb{E}[Y \mid X] = X^\top \beta.
$$

Let $x \in \mathbb{R}^n$ be a new input. So, we estimate

$$
\hat{\mathbb{E}}[Y \mid X = x] = x^\top \hat{\beta},
$$

where $\hat{\beta}$ is the least-squares estimate of β. We wish to estimate its class $y = f^*\!*(x)$, where $f^*\!*$ is the target function corresponding to 0–1 loss.

1. Derive the linear model estimate of $\hat{P}(Y = 1 \mid X = x)$.

2. Show that
$$
\hat{y} = \hat{f}^*(x) = 2 \cdot \mathbf{1}\{x^\top \hat{\beta} \geq 0\} - 1,
$$

where $\hat{f}^*$ is the estimate of $f^*\!*$ given by plugging in the estimated conditional probabilities $\hat{P}(Y = y \mid X = x)$.

Answers:

Setup: $\mathcal{X} = \mathbb{R}^n$, $\mathcal{Y} = \{-1, 1\}$. We model the conditional expectation as $E[Y|\mathbf{x}] = \mathbf{x}^T\beta$.
(a) Derive the linear model estimate of $\hat{P}(Y = 1|\mathbf{x})$.

$$
\begin{aligned}
\mathbb{E}[Y|\mathbf{x}] &= \sum_{y \in \{-1,1\}} y \cdot P(Y = y|\mathbf{x}) \\
&= (1)P(Y = 1|\mathbf{x}) + (-1)P(Y = -1|\mathbf{x}) \\
&= P(Y = 1|\mathbf{x}) - P(Y = -1|\mathbf{x})
\end{aligned}
$$

Since $P(Y = -1|\mathbf{x}) = 1 - P(Y = 1|\mathbf{x})$,

$$
\begin{aligned}
\mathbb{E}[Y|\mathbf{x}] &= P(Y = 1|\mathbf{x}) - (1 - P(Y = 1|\mathbf{x})) \\
&= 2P(Y = 1|\mathbf{x}) - 1
\end{aligned}
$$

Hence, solving for the probability:

$$
P(Y = 1|\mathbf{x}) = \frac{\mathbb{E}[Y|\mathbf{x}] + 1}{2}
$$

Plugging in our linear model estimate $\hat{\mathbb{E}}[Y|\mathbf{x}] = \mathbf{x}^T \hat{\beta}$:

$$\hat{P}(Y = 1|\mathbf{x}) = \frac{\mathbf{x}^T \hat{\beta} + 1}{2}$$

(b) Show that $\hat{y} = \hat{f}^*(\mathbf{x})$ is given by $2 \cdot \mathbb{1}_{\{\mathbf{x}^T \hat{\beta} \geq 0\}} - 1$. We know that the Bayes classifier $\hat{f}^*(\mathbf{x})$ picks the class with the highest estimated probability. Thus, we predict $\hat{y} = 1$ if $\hat{P}(Y = 1|\mathbf{x}) \geq \hat{P}(Y = -1|\mathbf{x})$.

$$\frac{\mathbf{x}^T \hat{\beta} + 1}{2} \geq 1 - \frac{\mathbf{x}^T \hat{\beta} + 1}{2}$$

$$\frac{\mathbf{x}^T \hat{\beta} + 1}{2} \geq \frac{1 - \mathbf{x}^T \hat{\beta}}{2}$$

$$\mathbf{x}^T \hat{\beta} + 1 \geq 1 - \mathbf{x}^T \hat{\beta}$$

$$2\mathbf{x}^T \hat{\beta} \geq 0 \implies \mathbf{x}^T \hat{\beta} \geq 0$$

This tells us that we predict class 1 for positive values of $\mathbf{x}^T \hat{\beta}$ and class -1 for negative values. This is indeed given by the function $2 \cdot \mathbb{1}_{\{\mathbf{x}^T \hat{\beta} \geq 0\}} - 1$:

- If $\mathbf{x}^T \hat{\beta} \geq 0$, then $2(1) - 1 = 1$.

- If $\mathbf{x}^T \hat{\beta} < 0$, then $2(0) - 1 = -1$.

2 Linear Regression

References

[1] D. S. Judson, *Abstract Algebra: Theory and Applications*, 3rd ed. Orthogonal Publishing L3C, 2019.