Deep learning

2.1. Loss and risk

François Fleuret

https://fleuret.org/dlc/

UNIVERSITÉ DE GENÈVE

The general objective of machine learning is to capture regularity in data to make predictions.

In our regression example, we modeled age and blood pressure as being linearly related, to predict the latter from the former.

The general objective of machine learning is to capture regularity in data to make predictions.

In our regression example, we modeled age and blood pressure as being linearly related, to predict the latter from the former.

There are multiple types of inference that we can roughly split into three categories:

- Classification (e.g. object recognition, cancer detection, speech processing),
- regression (e.g. customer satisfaction, stock prediction, epidemiology), and
- density estimation (e.g. outlier detection, data visualization, sampling/synthesis).

The standard formalization for classification and regression considers a measure of probability

$$\mu_{X,Y}$$

over the observation/value of interest, and i.i.d. training samples

$$(x_n, y_n), \; n = 1, \dots, N,$$

and for density estimation

$$\mu_X$$

and

$$x_n, \; n = 1, \dots N.$$

Intuitively, for classification a often intuitive interpretation is

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

that is, draw Y first, and given its value, generate X.

Intuitively, for classification a often intuitive interpretation is

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

that is, draw Y first, and given its value, generate X.

So the conditional distribution

$$\mu_{X|Y=y}$$

stands for the distribution of the observable signal for the class y (e.g. "sound of an /ē/", "image of a cat").

For regression, one would interpret the joint law more naturally as

$$\mu_{X,Y}(x,y) = \mu_{Y|X=x}(y)\,\mu_X(x)$$

which would be: first, generate X, and given its value, generate Y.

For regression, one would interpret the joint law more naturally as

$$\mu_{X,Y}(x, y) = \mu_{Y|X=x}(y)\, \mu_X(x)$$

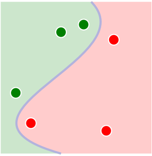
which would be: first, generate X, and given its value, generate Y.

In the simple cases

$$Y = f(X) + \epsilon$$

where f is the deterministic dependency between x and y (e.g. affine), and ϵ is a random noise, independent of X (e.g. Gaussian).

With such a probabilistic perspective, we can more precisely define the three types of inferences we introduced before:

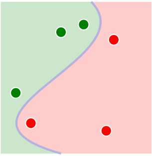


Classification,

- (X, Y) random variables on $\mathscr{Z} = \mathbb{R}^D \times \{1, \dots, C\}$,
- we want to estimate $\text{argmax}_y\, P(Y = y \mid X = x)$.

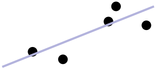
With such a probabilistic perspective, we can more precisely define the three types of inferences we introduced before:

Classification,



- (X, Y) random variables on $\mathscr{Z} = \mathbb{R}^D \times \{1, \dots, C\}$,
- we want to estimate $\operatorname{argmax}_y P(Y = y \mid X = x)$.

Regression,

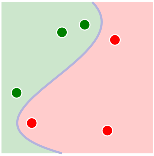


- (X, Y) random variables on $\mathscr{Z} = \mathbb{R}^D \times \mathbb{R}$,
- we want to estimate $\mathbb{E}(Y \mid X = x)$.

With such a probabilistic perspective, we can more precisely define the three types of inferences we introduced before:

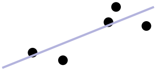
Classification,

- (X, Y) random variables on $\mathscr{Z} = \mathbb{R}^D \times \{1, \dots, C\}$,
- we want to estimate $\operatorname{argmax}_y P(Y = y \mid X = x)$.



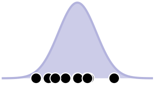
Regression,

- (X, Y) random variables on $\mathscr{Z} = \mathbb{R}^D \times \mathbb{R}$,
- we want to estimate $\mathbb{E}(Y \mid X = x)$.



Density estimation,

- X random variable on $\mathscr{Z} = \mathbb{R}^D$,
- we want to estimate μ_X.



The boundaries between these categories are fuzzy:

- Regression allows to do classification through class scores.
- Density models allow to do classification thanks to Bayes' law.

etc.

Risk, empirical risk

Learning consists of finding in a set $\mathscr{F}$ of functionals a "good" f^* (or its parameters' values) usually defined through a loss

$$\ell : \mathscr{F} \times \mathscr{Z} \to \mathbb{R}$$

such that $\ell(f, z)$ increases with how wrong f is on z.

Learning consists of finding in a set $\mathscr{F}$ of functionals a "good" f^* (or its parameters' values) usually defined through a loss

$$\ell : \mathscr{F} \times \mathscr{Z} \to \mathbb{R}$$

such that $\ell(f, z)$ increases with how wrong f is on z. For instance

- for classification:
$$\ell(f, (x, y)) = \mathbf{1}_{\{f(x) \neq y\}},$$
- for regression:
$$\ell(f, (x, y)) = (f(x) - y)^2,$$
- for density estimation:
$$\ell(q, z) = -\log q(z).$$

Learning consists of finding in a set $\mathscr{F}$ of functionals a "good" f^* (or its parameters' values) usually defined through a loss

$$\ell : \mathscr{F} \times \mathscr{Z} \rightarrow \mathbb{R}$$

such that $\ell(f, z)$ increases with how wrong f is on z. For instance

- for classification:

$$\ell(f, (x, y)) = \mathbf{1}_{\{f(x) \neq y\}},$$

- for regression:

$$\ell(f, (x, y)) = (f(x) - y)^2,$$

- for density estimation:

$$\ell(q, z) = -\log q(z).$$

The loss may include additional terms related to f itself.

We are looking for an f with a small **expected risk**

$$R(f) = \mathbb{E}_Z\left(\ell(f, Z)\right),$$

which means that our learning procedure would ideally choose

$$f^* = \underset{f \in \mathscr{F}}{\operatorname{argmin}}\, R(f).$$

We are looking for an f with a small **expected risk**

$$R(f) = \mathbb{E}_Z\left(\ell(f, Z)\right),$$

which means that our learning procedure would ideally choose

$$f^* = \underset{f \in \mathscr{F}}{\operatorname{argmin}}\, R(f).$$

Although this quantity is unknown, if we have i.i.d. training samples

$$\mathscr{D} = \left\{Z_1, \dots, Z_N\right\},$$

we can compute an estimate, the **empirical risk**:

$$\hat{R}(f; \mathscr{D}) = \hat{\mathbb{E}}_{\mathscr{D}}(\ell(f, Z)) = \frac{1}{N} \sum_{n=1}^{N} \ell(f, Z_n).$$

We have

$$\mathbb{E}_{Z_1,\dots,Z_N}\left(\hat{R}(f;\mathscr{D})\right) \quad = \quad \mathbb{E}_{Z_1,\dots,Z_N}\left(\frac{1}{N}\sum_{n=1}^{N}\ell(f,Z_n)\right)$$

We have

$$
\begin{aligned}
\mathbb{E}_{Z_1,\dots,Z_N}\left(\hat{R}(f;\mathscr{D})\right) &= \mathbb{E}_{Z_1,\dots,Z_N}\left(\frac{1}{N}\sum_{n=1}^{N}\ell(f,Z_n)\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{Z_n}\left(\ell(f,Z_n)\right)
\end{aligned}
$$

We have

$$\begin{aligned}\mathbb{E}_{Z_1,\dots,Z_N}\left(\hat{R}(f;\mathscr{D})\right) &= \mathbb{E}_{Z_1,\dots,Z_N}\left(\frac{1}{N}\sum_{n=1}^{N}\ell(f,Z_n)\right)\\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{Z_n}\left(\ell(f,Z_n)\right)\\ &= \mathbb{E}_{Z}\left(\ell(f,Z)\right)\end{aligned}$$

We have

$$\begin{aligned}
\mathbb{E}_{Z_1,\dots,Z_N}\left(\hat{R}(f;\mathscr{D})\right) &= \mathbb{E}_{Z_1,\dots,Z_N}\left(\frac{1}{N}\sum_{n=1}^{N}\ell(f,Z_n)\right)\\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{Z_n}\left(\ell(f,Z_n)\right)\\
&= \mathbb{E}_{Z}\left(\ell(f,Z)\right)\\
&= R(f).
\end{aligned}$$

We have

$$\begin{aligned}
\mathbb{E}_{Z_1,\dots,Z_N}\left(\hat{R}(f;\mathscr{D})\right) &= \mathbb{E}_{Z_1,\dots,Z_N}\left(\frac{1}{N}\sum_{n=1}^{N}\ell(f,Z_n)\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{Z_n}\left(\ell(f,Z_n)\right) \\
&= \mathbb{E}_{Z}\left(\ell(f,Z)\right) \\
&= R(f).
\end{aligned}$$

The empirical risk is an **unbiased estimator** of the expected risk.

Finally, given $\mathscr{D}$, $\mathscr{F}$, and ℓ, "learning" aims at computing

$$f^* = \operatorname*{argmin}_{f \in \mathscr{F}} \hat{R}(f; \mathscr{D}).$$

Finally, given $\mathscr{D}$, $\mathscr{F}$, and ℓ, "learning" aims at computing

$$f^* = \underset{f \in \mathscr{F}}{\operatorname{argmin}} \hat{R}(f; \mathscr{D}).$$

- Can we bound $R(f)$ with $\hat{R}(f; \mathscr{D})$?

  Yes if f is not chosen using $\mathscr{D}$. Since the Z_n are independent, we just need to take into account the variance of $\hat{R}(f; \mathscr{D})$.

Finally, given $\mathscr{D}$, $\mathscr{F}$, and ℓ, "learning" aims at computing

$$f^* = \underset{f \in \mathscr{F}}{\operatorname{argmin}} \hat{R}(f; \mathscr{D}).$$

- Can we bound $R(f)$ with $\hat{R}(f; \mathscr{D})$?

  Yes if f is not chosen using $\mathscr{D}$. Since the Z_n are independent, we just need to take into account the variance of $\hat{R}(f; \mathscr{D})$.

- Can we bound $R(f^*)$ with $\hat{R}(f^*; \mathscr{D})$?

  Unfortunately not simply, and not without additional constraints on $\mathscr{F}$.

Finally, given $\mathscr{D}$, $\mathscr{F}$, and ℓ, "learning" aims at computing

$$f^* = \underset{f \in \mathscr{F}}{\operatorname{argmin}}\, \hat{R}(f; \mathscr{D}).$$

- Can we bound $R(f)$ with $\hat{R}(f; \mathscr{D})$?

  Yes if f is not chosen using $\mathscr{D}$. Since the Z_n are independent, we just need to take into account the variance of $\hat{R}(f; \mathscr{D})$.

- Can we bound $R(f^*)$ with $\hat{R}(f^*; \mathscr{D})$?

  Unfortunately not simply, and not without additional constraints on $\mathscr{F}$.

  For instance if $|\mathscr{F}| = 1$, we can!

Note that in practice, we call "loss" both the functional

$$\ell : \mathscr{F} \times \mathscr{Z} \to \mathbb{R}$$

and the empirical risk minimized during training

$$\mathscr{L}(f) = \frac{1}{N} \sum_{n=1}^{N} \ell(f, z_n).$$

The end