Basic "Supervised" Machine Learning Setup

Given, i.e. exists indep. of what we do.

- Input Space $\mathcal{X}$
- Output Space $\mathcal{Y}$
- Unknown Distribution μ on $\mathcal{X} \times \mathcal{Y}$.
  - We are able to sample from $\mathcal{X} \times \mathcal{Y}$ in some way.

Contains our modelling assum- and objects.

We provide:

- A class of functions (parametrized by Θ), $f(\cdot, \Theta) : \mathcal{X} \to \mathcal{Y}'$.
- A loss function $L : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}_{\geq 0}$.

Goal:

- Minimize the Risk:

$$\Theta^* = \operatorname{argmin}_{\Theta} \int_{\mathcal{X} \times \mathcal{Y}} L(y, f(x, \Theta)) \, d\mu .$$

Examples:

- Regression (L^2-regression):
  - Let $\mathcal{X} = [0,1]$, $\mathcal{Y} = \mathbb{R}$, and μ given by the pushforward of the ... under the map $x \to (x, g(x))$ for some unknown function.
  - Let $f(\cdot, \Theta)$ be some class of function C (say linear or polynomial).
  - Let $L(y, y') = |y - y'|^2$

$$\to \operatorname{argmin}_{f \in C} \int_0^1 |f(x) - g(x)|^2 \, dx \; \} = R(f)$$

- Hard-Classification:
  - Let $\mathcal{X} = \mathbb{R}^n$, $\mathcal{Y} = \{\pm 1\}$
  - Let $l : \mathcal{X} \to \mathcal{Y}$ be an (unknown) labelling function, and μ a dist. on $\mathcal{X}$.
  - Let $f(\cdot, \Theta)$ be
  - $L(y, y') = \mathbb{1}_{y \neq y'}$

→ $R(f) = \mathbb{P}_\mu(f(x) \neq \ell(x))$

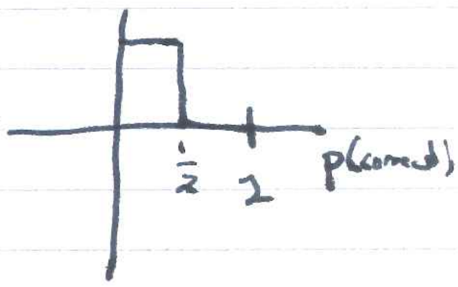
- Modified Classification

⋮

· $Y' = \mathbb{P}(Y = \{\pm 1\})$

so $f(\cdot, \theta): X \to Y'$

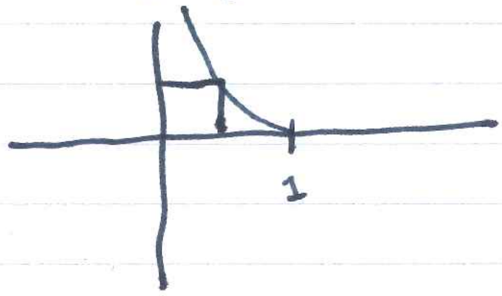
· $L(y, y') = \begin{cases} 1 & \text{if } p_{y'}(y) \leq \frac{1}{2} \\ 0 & \text{"} \geq \frac{1}{2} \end{cases}$



- Cross Entropy for classification

"
- k
"

· $L(y, y') = -\log_2(p_{y'}(y))$



Note: Cross-entropy is an upper bound on class. loss.

· We additionally get penalized for our (lack of) confidence.

⋮

Empirical Risk

· We don't have access to μ, but we can draw samples $(x_1, y_1), \ldots, (x_n, y_n)$

$$G^TG - \frac{\sqrt{2}}{2}(G^TY)^2 - \left(1-\frac{\sqrt{2}}{2}\right)(G^TY)^2$$
$$-\frac{\sqrt{2}}{2}(Y^TG)^2 - \left(1-\frac{\sqrt{2}}{2}\right)(Y^TG)^2 + \frac{1}{2}Y^TGG^TY$$

Consider instead the (regularized) empirical loss:

$$+\left(1-\frac{\sqrt{2}}{2}\right)^2 Y^TGG^TY + \frac{\sqrt{2}}{2}\left(1-\frac{\sqrt{2}}{2}\right)\left[(Y^TG)^2 + (G^TY)^2\right]$$

$$R^n_{emp}(\theta) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i,\theta)) \; (+\lambda S(\theta))$$

$$2 + \frac{1}{2} + 1 - \sqrt{2} + \frac{1}{2} - \sqrt{2} + 1 \overset{?}{=} 4 - 2\sqrt{2}$$

and consider $\theta^*_{emp,n} = \underset{\theta}{\operatorname{argmin}}\, R^n_{emp}(\theta)$

Central question:

- Does
$$R^n_{emp}(\theta^*_{emp,n}) \xrightarrow{?} R(\theta^*)$$

- Note that the weak law of large numbers means that for a fixed θ,
$$R^n_{emp}(\theta) \xrightarrow{prob} R(\theta)$$

i.e. $$\mathbb{P}(|R^n_{emp}(\theta) - R(\theta)| > \varepsilon) \to 0 \quad \forall \varepsilon > 0.$$

- However, it may be that the convergence isn't uniform! This is the phenomenon of overfitting.

$$V^TX + X^TV \qquad \left(I - \frac{1}{2}XX^T\right) = A$$
$$V = A^{-1/2}W$$
$$V' = A^{1/2}W' \qquad W^TA^{-1/2}X + X^TA^{-1/2}W = 0$$

$$W' = A^{-1/2}V' - \frac{1}{2}\left(V'^TA^{-1/2}X + X^TA^{-1/2}V'\right)$$

X

$$A^{1/2}W' = V' - \frac{1}{2}X\left(V'^TA^{-1/2}X + X^TA^{-1/2}V'\right)$$

$A^{1/2}$

$$A^{-1} = (I + XX^T)$$

Overfitting Example:

- Consider the first example of last class:
$X = [0,1]$, $Y = \mathbb{R}$, μ on $X \times Y$ given by sampling X uniformly and setting $y = g(x)$ for some unknown function g.
- Let $f(\cdot, \theta)$ be the set of polynomials, i.e. $\Theta = \{ (a_i)_{i=0}^{\infty} \in \mathbb{R}^{\mathbb{N}}$ s.t. $\exists N$, s.t. $a_i = 0, i \geq N \}$ $\quad Y' = \mathbb{R}$

$$f(x, \theta) = \sum_{i=0}^{\infty} a_i x^i =: p_\theta(x)$$

$\cdot\ L(y, y') = |y - y'|^2$

- The risk is

$$R(\theta) = \int |p_\theta(x) - g(x)|^2 \, dx.$$

The empirical risk is

$$R^n_{emp}(\theta) = \frac{1}{n} \sum_{i=1}^{n} (p_\theta(x_i) - g(x_i))^2.$$

Let $\theta^*_n \in \operatorname{argmin} R^n_{emp}$ be the degree $n-1$ interpolation at the points $x_1, \dots, x_n$, then

$$R^n_{emp}(\theta^*_n) = 0.$$

Does $R(\theta^*_n) \to 0$? No! Not ~~in general~~ for general g.

Problem: For each fixed θ,

$R^n_{emp}(\theta) \longrightarrow R(\theta)$, but the minimizer θ^*_n of R^n_{emp} is always such that this convergence is especially slow! Need uniform convergence!

Note:

- If g is smooth enough: No overfitting
- If the points x_i are sampled differently (see Chebyshev points / Chebyshev nodes): Also no overfitting

$\rightarrow$ Overfitting depends both on the model and the distribution μ on $X \times Y$.