RegML 2016
Class 1
Statistical Learning Theory

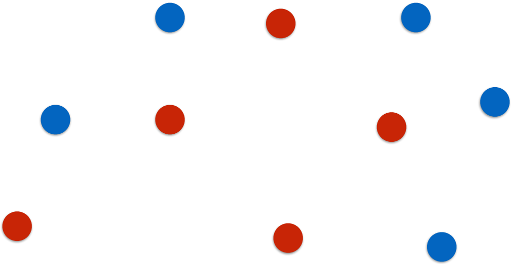
Lorenzo Rosasco
UNIGE-MIT-IIT

June 27, 2016

All starts with DATA

- **Supervised:** $\{(x_1, y_1), \ldots, (x_n, y_n)\}$,
- Unsupervised: $\{x_1, \ldots, x_m\}$,
- Semi-supervised: $\{(x_1, y_1), \ldots, (x_n, y_n)\} \cup \{x_1, \ldots, x_m\}$

Learning from examples



Problem: **given** $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ **find** $f(x_{\text{new}}) \sim y_{\text{new}}$

Setting for the supervised learning problem

- $X \times Y$ probability space, with measure ρ.
- $S_n = (x_1, y_1), \ldots, (x_n, y_n) \sim \rho^n$, i.e. sampled i.i.d.
- $L : Y \times Y \to [0, \infty)$, measurable *loss function*.
- Expected risk

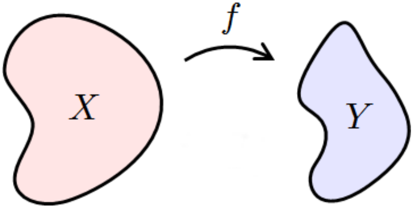
$$\mathcal{E}(f) = \int_{X \times Y} L(y, f(x)) d\rho(x, y).$$

Problem: Solve

$$\min_{f : X \to Y} \mathcal{E}(f),$$

given only S_n (ρ fixed, but unknown).

Data space



$$\underbrace{X}_{\text{input space}} \qquad \underbrace{Y}_{\text{output space}}$$

Input space

X input space:

- linear spaces, e. g.
  - vectors,
  - functions,
  - matrices/operators
- "structured" spaces, e. g.
  - strings,
  - probability distributions,
  - graphs

Output space

Y output space

- linear spaces, e. g.
  - $Y = \mathbb{R}$, regression,
  - $Y = \{+1, -1\}$, classification,
  - $Y = \mathbb{R}^T$, multi-task regression,
  - $Y = \{1, \ldots, T\}$, multi-label classification
- *"structured" spaces*
  - strings,
  - probability distributions,
  - graphs

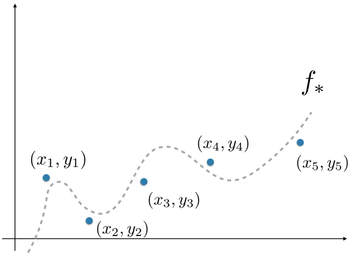
Probability distribution

Reflects *uncertainty* and *stochasticity* of the learning problem

$$\rho(x, y) = \rho_X(x)\rho(y|x),$$

- ρ_X marginal distribution on X,
- $\rho(y|x)$ conditional distribution on Y given $x \in X$.

Conditional distribution and noise



Regression

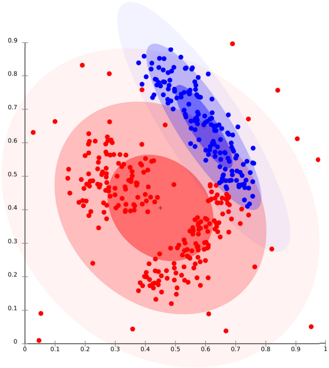
$$y_i = f_*(x_i) + \epsilon_i,$$

- Let $f_* : X \to Y$, fixed function
- $\epsilon_1, \dots, \epsilon_n$ zero mean random variables
- $x_1, \dots, x_n$ random

Conditional distribution and misclassification

Classification

$$\rho(y|x) = \{\rho(1|x), \rho(-1|x)\},$$

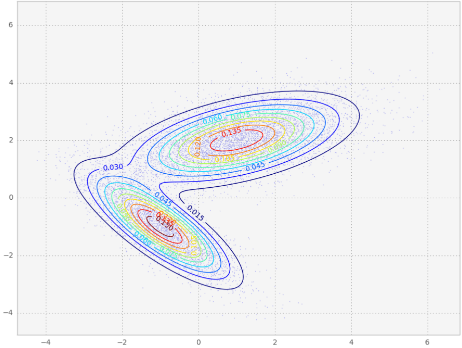


Noise in classification: overlap between the classes

$$\Delta_t = \left\{ x \in X \;\middle|\; \left|\rho(1|x) - \rho(-1|x)\right| \leq t \right\}$$

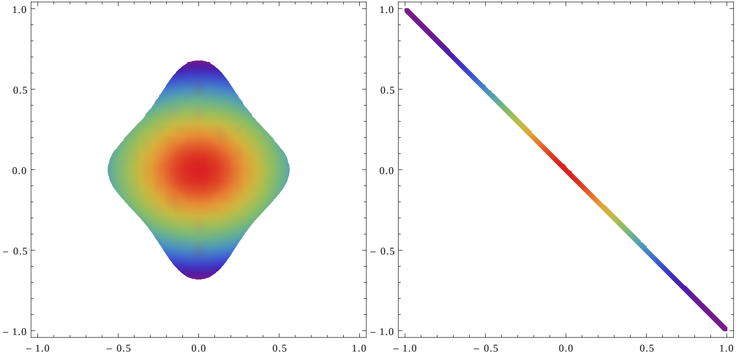
Marginal distribution and sampling

ρ_X takes into account uneven sampling of the input space



Marginal distribution, densities and manifolds

$$p(x) = \frac{d\rho_X(x)}{dx} \to p(x) = \frac{d\rho_X(x)}{d\text{vol}(x)},$$



Loss functions

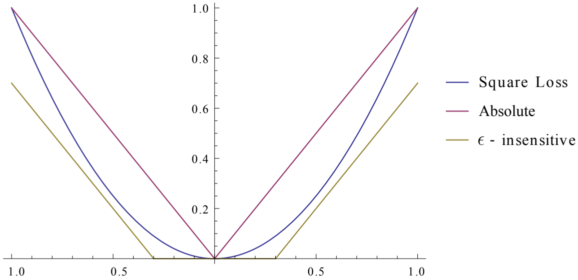
$$L : Y \times Y \to [0, \infty),$$

- The cost of predicting $f(x)$ in place of y.
- Part of the problem definition $\mathcal{E}(f) = \int L(y, f(x)) d\rho(x, y)$
- Measures the *pointwise error*,

Losses for regression

$$L(y, y') = L(y - y')$$

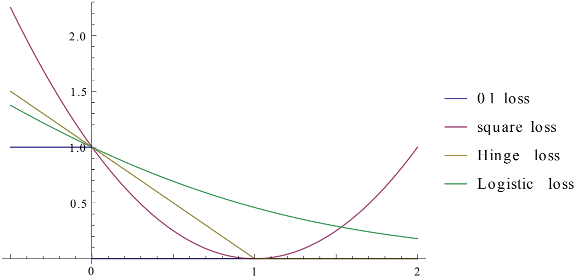
- Square loss $L(y, y') = (y - y')^2$,
- Absolute loss $L(y, y') = |y - y'|$,
- ϵ-insensitive $L(y, y') = \max(|y - y'| - \epsilon, 0)$,



Losses for classification

$$L(y, y') = L(-yy')$$

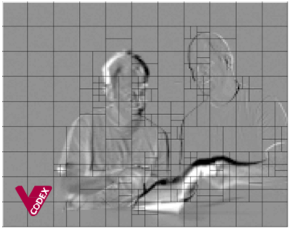
- 0-1 loss $L(y, y') = \mathbf{1}_{\{-yy'>0\}}$
- Square loss $L(y, y') = (1 - yy')^2$,
- Hinge-loss $L(y, y') = \max(1 - yy', 0)$,
- logistic loss $L(y, y') = \log(1 + \exp(-yy'))$,



Losses for structured prediction

Loss specific for each learning task e. g.

- Multi-class: square loss, weighted square loss, logistic loss, . . .
- Multi-task: weighted square loss, absolute, . . .
- . . .



Expected risk

$$\mathcal{E}(f) = \mathcal{E}_L(f) = \int_{X \times Y} L(y, f(x)) d\rho(x, y)$$

note that $f \in \mathcal{F}$ where

$$\mathcal{F} = \{f : X \to Y \mid f \text{ measurable}\}.$$

Example $Y = \{-1, +1\}, \quad L(y, f(x)) = \mathbf{1}_{\{-yf(x) > 0\}}$

$$\mathcal{E}(f) = \mathbb{P}(\{(x, y) \in X \times Y \mid f(x) \neq y\}).$$

Target function

$$f_\rho \quad = \quad \arg\min_{f \in \mathcal{F}} \ \mathcal{E}(f),$$

can be derived for many loss functions...

Target functions in regression

square loss,

$$f_\rho(x) = \int_Y y d\rho(y|x)$$

absolute loss,

$$f_\rho(x) = \text{median } \rho(y|x),$$

where

$$\text{median } p(\cdot) = y \ \text{ s.t. } \int_{-\infty}^{y} t dp(t) = \int_{y}^{+\infty} t dp(t).$$

Target functions in classification

0-1 loss,

$$f_\rho(x) = \mathbf{sign}(\rho(1|x) - \rho(-1|x))$$

square loss,

$$f_\rho(x) = \rho(1|x) - \rho(-1|x)$$

logistic loss,

$$f_\rho(x) = \log \frac{\rho(1|x)}{\rho(-1|x)}$$

hinge-loss,

$$f_\rho(x) = \mathbf{sign}(\rho(1|x) - \rho(-1|x))$$

Learning algorithms

$$S_n \to \widehat{f}_n = \widehat{f}_{S_n}$$

f_n estimates f_ρ given the observed examples S_n

How to measure the error of an estimator?

Excess risk

Excess Risk:

$$\mathcal{E}(\widehat{f}) - \inf_{f \in \mathcal{F}} \mathcal{E}(f),$$

Consistency: For any $\epsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}\left(\mathcal{E}(\widehat{f}) - \inf_{f \in \mathcal{F}} \mathcal{E}(f) > \epsilon\right) = 0,$$

Tail bounds, sample complexity and error bound

- *Tail bounds*: For any $\epsilon > 0, n \in \mathbb{N}$

$$\mathbb{P}\left(\mathcal{E}(\widehat{f}) - \inf_{f \in \mathcal{F}} \mathcal{E}(f) > \epsilon\right) \leq \delta(n, \mathcal{F}, \epsilon)$$

- Sample complexity: For any $\epsilon > 0, \delta \in (0, 1]$, when $n \geq n_0(\epsilon, \delta, \mathcal{F})$

$$\mathbb{P}\left(\mathcal{E}(\widehat{f}) - \inf_{f \in \mathcal{F}} \mathcal{E}(f) > \epsilon\right) \leq \delta,$$

- *Error bounds*: For any $\delta \in (0, 1]$, $n \in \mathbb{N}$, with probability at least $1 - \delta$,

$$\mathcal{E}(\widehat{f}) - \inf_{f \in \mathcal{F}} \mathcal{E}(f) \leq \epsilon(n, \mathcal{F}, \delta),$$

Error bounds and no free-lunch theorem

Theorem For any $\widehat{f}$, there exists a problem for which

$$\mathbb{E}(\mathcal{E}(\widehat{f}) - \inf_{f \in \mathcal{F}} \mathcal{E}(f)) > 0$$

No free-lunch theorem continued

Theorem For any $\widehat{f}$, there exists a ρ such that

$$\mathbb{E}(\mathcal{E}(\widehat{f}) - \inf_{f \in \mathcal{F}} \mathcal{E}(f)) > 0$$

$\mathcal{F} \to \mathcal{H}$ *Hypothesis space*

Hypothesis space

$$\mathcal{H} \subset \mathcal{F}$$

E.g. $X = \mathbb{R}^d$

$$\mathcal{H} = \{f(x) = \langle w, x \rangle = \sum_{j=1}^{d} w_j x_j, \mid w \in \mathbb{R}^d, \forall x \in X\}$$

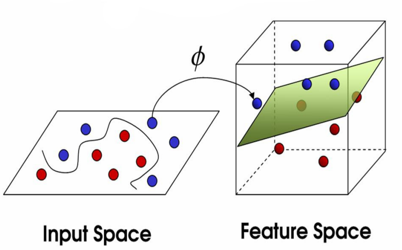
then $\mathcal{H} \simeq \mathbb{R}^d$.

Finite dictionaries

$$D = \{\phi_i : X \to \mathbb{R} \mid i = 1, \dots, p\}$$

$$\mathcal{H} = \{f(x) = \sum_{j=1}^{p} w_j \phi_j(x) \mid w_1, \dots, w_p \in \mathbb{R}, \forall x \in X\}$$

$$f(x) = w^\top \Phi(x), \quad \Phi(x) = (\phi_1(x), \dots, \phi_p(x))$$



This class

Learning theory ingredients

- Data space/distribution
- Loss function, risks and target functions
- Learning algorithms and error estimates
- Hypothesis space

Next class

- Regularized learning algorithm: penalization
- Statistics and computations
- Nonparametrics and kernels