Introduction to Supervised Learning

B. Michel

Ecole Centrale de Nantes

The supervised learning problem

- Outcome measurement Y (also called dependent variable, response, target).
- Input : vector of p measurements X (also called regressors, covariates, features).
- In the regression problem, $Y \in \mathbb{R}$ is quantitative (e.g. price).
- In the classification problem, Y takes values in a finite, unordered set (e.g. digit 0-9, image classification).
  Binary classification : $Y \in \{0, 1\}$ or $\{-1, 1\}$.
- Structured prediction : Y is a more complex quantity : vector (multivariate regression) graph, tree, string, image ...

The supervised learning problem

- We have training data $\mathcal{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$.
- Probabilistic framework: (X_i, Y_i) : i.i.d. replications of some random variable $(X, Y) \in \mathcal{X} \times \mathcal{Y}$.
- $\mathcal{X}$ and $\mathcal{Y}$ are assumed to be measurable sets.
- Let P be the joint probability of (X, Y).
- We want to infer the link between the input variables X and the outcome Y.
- A predictor is a function $f : \mathcal{X} \to \mathcal{Y}$.
- Our goal : use $\mathcal{D}_n$ to build up a predictor $\hat{f}$ such that $\hat{f}(X) \approx Y$.

Statistical Learning in practice (regression / classification)

1. Data Collection
2. Feature engineering (sampling, feature extraction, feature transformation ...)
3. Choose a statistical model (parametric or not) or an algorithm to find a learning rule $\hat{f}$
4. **Fit** (statistics and optimization) the learning rule $\hat{f}$ or a collection of learning rules $(\hat{f})_{\lambda \in \Lambda}$.
5. Tune the parameters (typically by cross validation)
6. Make prediction with the final model.

Generalization error

- We consider a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ s.t. $\ell(y, y) = 0$ and $\ell(y, y') \geq 0$.
- **Risk** of a predictor $f : \mathcal{X} \to \mathcal{Y}$, also called **generalization error** or **prediction error** :

$$\mathcal{R}(f) = \mathbb{E}\ell(f(X), Y)$$

where the expectation is under P.

Generalization error

- We consider a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ s.t. $\ell(y, y) = 0$ and $\ell(y, y') \geq 0$.

- **Risk** of a predictor $f : \mathcal{X} \to \mathcal{Y}$, also called **generalization error** or **prediction error** :

$$\mathcal{R}(f) = \mathbb{E}\ell(f(X), Y)$$

  where the expectation is under P.

- Loss functions for regression :
  - Quadratic loss : $\ell(a, b) = (a - b)^2$.
  - Absolute loss : $\ell(a, b) = |a - b|$.

Generalization error

- We consider a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ s.t. $\ell(y, y) = 0$ and $\ell(y, y') \geq 0$.

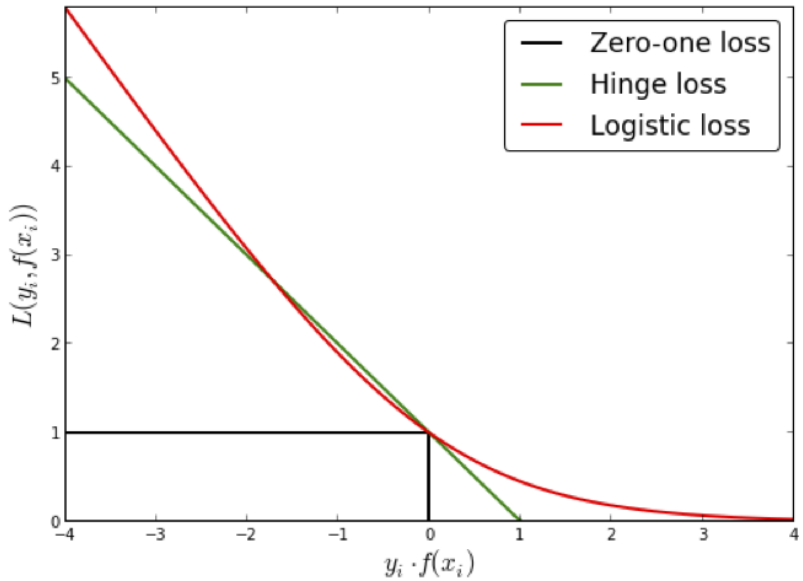
- **Risk** of a predictor $f : \mathcal{X} \to \mathcal{Y}$, also called **generalization error** or **prediction error** :

$$\mathcal{R}(f) = \mathbb{E}\ell(f(X), Y)$$

  where the expectation is under P.

- Loss functions for Classification :

  - 0-1 loss :
    $\ell(a, b) = \mathbb{1}_{a \neq b}$.
  - Hinge loss :
    $\ell(a, b) = \max(0, 1 - ab)$
  - Logistic loss :
    $\ell(a, b) = \log[1 + \exp(-ab)]$



Bayes risk

- Optimal risk or *Bayes risk*

$$\mathcal{R}^\star = \inf_{f \in \mathcal{F}} \mathcal{R}(f)$$

  where $\mathcal{F}$ is the set of all possible predictors.
- $\mathcal{R}^\star$ depends on ℓ and P.
- Optimal predictor or Bayes predictor $f^\star$:

$$\mathcal{R}(f^\star) = \mathcal{R}^\star$$

- In most cases $\mathcal{R}(f^\star) > 0$.
- Bayes predictor : not unique and does not always exists.

Learning rule

- We use the observations $\mathcal{D}_n$ to find a **learning rule** (or decision rule) $\hat{f}$ s.t. $\mathcal{R}(\hat{f})$ is as small as possible.
- The risk is conditional to $\mathcal{D}_n$:

$$\mathcal{R}(\hat{f}) = \mathbb{E}\left[\ell(\hat{f}(X), Y)|\mathcal{D}_n\right]$$

- We also consider the mean risk :

$$\mathbb{E}\mathcal{R}(\hat{f}) = \mathbb{E}\left[\ell(\hat{f}(X), Y)\right].$$

Regression

- Here we assume that $\mathcal{Y} = \mathbb{R}$.
- $\mathcal{Y} = \mathbb{R}^d$ corresponds to multivariate regression (régression multivariée).
- Learning rules for regression :
  - Linear predictors
  - Decision trees
  - neural networks
  - ...

Regression

- Here we assume that $\mathcal{Y} = \mathbb{R}$.
- $\mathcal{Y} = \mathbb{R}^d$ corresponds to multivariate regression (régression multivariée).
- Learning rules for regression :
  - Linear predictors
  - Decision trees
  - neural networks
  - ...
- If $\mathbb{E}|Y| < \infty$, we introduce the **regression function** $\eta : \mathcal{X} \to \mathbb{R}$:

$$\eta(X) := \mathbb{E}[Y|X] \quad \text{a.s.}$$

and the error $\varepsilon = Y - \eta(X)$. Finally :

$$Y = \eta(X) + \varepsilon$$

where $\mathbb{E}[\varepsilon|X] = 0$ a.s.

Regression

- For the quadratic loss $\ell(a, b) = (b - a)^2$, the risk $\mathcal{R}(f)$ is called the quadratic risk.
- The regression function η is "the" best predictor:

Proposition (Quadratic regression)

Assume that $\mathcal{Y} = \mathbb{R}$ and $\mathbb{E}(\varepsilon^2) < \infty$. Then the regression function η is a Bayes predictor $(= f^\star)$.

- Of course η is unknown in practice so we will have to use learning rules.

Binary classification

- We assume that $\mathcal{Y} = \{0, 1\}$
- Learning rules for Classification :
  - Linear classifiers
  - Decision trees
  - neural networks
  - ...

Binary classification

- We assume that $\mathcal{Y} = \{0, 1\}$
- Learning rules for Classification :
  - Linear classifiers
  - Decision trees
  - neural networks
  - ...
- We also define the regression function $\eta : \mathcal{X} \to \mathbb{R}$:

$$\eta(X) := \mathbb{E}[Y|X] = P(Y = 1|X)$$

- For the 0-1 loss $\ell(a, b) = \mathbb{1}_{a \neq b}$:

$$\mathcal{R}(f) = \mathbb{E}(\mathbb{1}_{Y \neq f(X)}) = P(f(X) \neq Y).$$

Proposition

The classifier $f(x) = \mathbb{1}_{\eta(x) > 1/2} = f^\star(x)$ is a Bayes classifier.

Empirical risk

- Data $\mathcal{D}_n = (X_1, Y_1), \ldots, (X_n, Y_n)$.
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$
- The risk $\mathcal{R}(f) = \mathbb{E}\left[\ell(f(X), Y)\right]$ of a predictor f is unknown.
- A natural approach : estimate $\mathcal{R}(f)$ with the empirical risk

$$\hat{\mathcal{R}}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

- $\hat{\mathcal{R}}_n(f)$ is an unbiased estimator of $\mathcal{R}(f)$.

Empirical Risk Minimization (ERM)

- S : a set of predictors (i.e. $S \subset \mathcal{F}$)
- We say that $\hat{f} \in S$ is a minimizer of the empirical risk over S if

$$\hat{f} \in \operatorname*{argmin}_{g \in S} \hat{\mathcal{R}}_n(g).$$

- ERM : no existence and uniqueness.
- It is not always possible to compute the ERM.
- Example: linear regression provides an ERM for the loss ℓ_2 in the space of predictors which can be written as linear combinations of the variables.

Discriminative vs Generative Methods

- **Generative Approaches (rather for classification)** :
  - Based on the **conditional distribution** $(X|Y)$ distribution and the Bayes Theorem.
  - "Generative" because it is based on the joint distribution (X, Y) that generates the observations.
  - Popular models : Gaussians, Naive Bayes, Linear / Quadratic Discriminant Analysis, Hidden Markov Models (HMM), Bayesian networks, Markov random fields ...

- **Discriminative Approaches** : Directly optimize over a class of predictor.
  - based on the conditional likelihood $(Y|X)$, and least squares approaches (or other) to fit the model
  - Popular models : Linear regression, Logistic regression, SVMs, neural networks, Nearest neighbour, Conditional Random Fields (CRF), Random Forests, boosting ...