Apprentissage supervisé

Guillaume Obozinski

Ecole des Ponts - ParisTech

Cours MALAP 2014

What kind of learning?

Learn to:

- Recognize different kinds of butterflies from specimens
- Detect pedestrians on the street with an on board camera
- Read postal codes/checks
- Produce the syntactical relations between words in a sentence
- Predict which chemical components can react with a given protein
- Translate from a language to another
- Recognize speech
- Fly a helicopter

Learn *empirically* from a flow of experience, i.e. from a data stream

Outline

Supervised learning

Supervised learning

Setting:
Data come in pairs (x, y) of

- x some input data, often a vector of numerical features or descriptors (stimuli)
- y some output data

Goal:
Given some examples of existing pairs (x_i, y_i), "guess" some of the statistical relation between x and y that are relevant to a task.

Formalizing supervised learning

We will assume that we have some **training data**

$$D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}.$$

Learning scheme or learning "algorithm"

- is a functional $\mathscr{A}$ which
- given some training data D_n
- produces a predictor or decision function $\widehat{f}$.

$$\begin{array}{rccc} \mathscr{A}: & \bigcup_{n \in \mathcal{N}} (\mathcal{X} \times \mathcal{Y})^n & \rightarrow & \mathcal{Y}^{\mathcal{X}} \\ & D_n & \mapsto & \widehat{f} \end{array}$$

We hope to get a "good" decision function

$\rightarrow$ Need to define what we expect from that decision function.

Decision theory

Abraham Wald (1939)

Decision theoretic framework

- $\mathcal{X}$ input data set
- $\mathcal{Y}$ output data set
- $\mathcal{A}$ action set
- $f : \mathcal{X} \to \mathcal{A}$ decision function, predictor, hypothesis

Goal of learning
Produce a decision function such that given a new input x the action $f(x)$ is a "good" action when confronted to the unseen corresponding output y.

What is a "good" action?

- $f(x)$ is a good prediction of y, i.e. close to y in some sense.
- $f(x)$ is action that has the smallest possible cost when y occurs.

Loss function

$$\begin{array}{rccc} \ell : & \mathcal{A} \times \mathcal{Y} & \to & \mathbb{R} \\ & (a, y) & \mapsto & \ell(a, y) \end{array}$$

measures the cost incurred when action a is taken and y has occurred.

Generalization and expected behavior

Minimize worst future cost *vs* average future cost?

- Given x there might be some intrinsic uncertainty about y.
- To *generalize* to new pairs (x, y) they have to be similar to what has been encountered in the past.
- The worst possible (x, y) might be too rare.

Assume that the data is generated by

- by a stationary stochastic process.
- as independent and identically distributed random variables (X_i, Y_i)

Formalizing the goal of learning as minimizing the risk

Risk

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

Target function

If there *exists* a *unique* function f^* such that $\mathcal{R}(f^*) = \inf_{f \in \mathcal{Y}^{\mathcal{X}}} \mathcal{R}(f)$, then f^* is called the *target function*, *oracle function* or *Bayes predictor*.

Conditional risk

$$\mathcal{R}(a \,|\, x) = \mathbb{E}[\ell(a, Y) \mid X = x] = \int \ell(a, y)\, dP_{Y|X}(y|x).$$

If $\inf_{a \in \mathcal{A}} \mathcal{R}(a \,|\, x)$ is attained and unique for almost all x then the function $f^*(x) = \arg\min_{a \in \mathcal{A}} \mathcal{R}(a \,|\, x)$ is the target function.

Excess risk

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

Example 1: ordinary least squares regression

Case where $\mathcal{A} = \mathcal{Y} = \mathbb{R}$.

- square loss:

$$\ell(a, y) = \frac{1}{2}(a - y)^2$$

- mean square risk:

$$\begin{aligned} \mathcal{R}(f) &= \frac{1}{2}\mathbb{E}\left[(f(X) - Y)^2\right] \\ &= \frac{1}{2}\mathbb{E}\left[(f(X) - \mathbb{E}[Y|X])^2\right] + \frac{1}{2}\mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2\right] \end{aligned}$$

- target function:

$$f^*(X) = \mathbb{E}[Y|X]$$

Example 2: classification

Case where $\mathcal{A} = \mathcal{Y} = \{0, \ldots, K-1\}$.

- 0-1 loss:

$$\ell(a, y) = 1_{\{a \neq y\}}$$

- the risk is the misclassification error

$$\mathcal{R}(f) = \mathbb{P}(f(X) \neq Y)$$

- the target function is the assignment to the most likely class

$$f^*(X) = \mathrm{argmax}_{1 \leq k \leq K} \mathbb{P}(Y = k | X)$$

Example 3: sequence decoding (OCR)

Given $X = (X_1, \ldots, X_m) \in \mathcal{X}$ predict $Y = (Y_1, \ldots, Y_m)$.

- input space $\mathcal{X} = \left(\mathbb{R}^p\right)^m$ and output space $\mathcal{Y} = \mathcal{A} = \mathcal{S}^m$
- predictors $f = (f_1, \ldots, f_m)$ with $f_i : \mathcal{X} \to \mathcal{S}$
- Hamming loss

$$\ell_H(y, a) = \sum_{j=1}^{m} 1_{\{a_j \neq y_j\}}$$

- Combined loss

$$\ell(a, y) = c_{0\text{-}1}\, 1_{\{a \neq y\}} + c_H\, \ell_H(y, a)$$

- Risk

$$c_{0\text{-}1}\, \mathbb{P}(Y \neq f(X)) + c_H \sum_{j=1}^{m} \mathbb{P}(Y_i \neq f_i(X))$$

Example 4: ranking pairs

Assume that given a pair of random variables $(X, X') \in \mathcal{X}^2$ a preference variable $Y \in \{-1, 1\}$ is defined. Learn a score function on the variable X which is higher for the preferred instances.

- input variables $(X, X') \in \mathcal{X}^2$ with same distribution
- output variable: $Y \in \mathcal{Y} = \{-1, 1\}$
- action space: $\mathbb{R}$
- predictor $f : X \mapsto f(X)$
- loss:

$$\ell(a, b, y) = 1_{\{(a-b)\, y \geq 0\}}$$

- risk:

$$\mathbb{P}\big(Y\big[f(X) - f(X')\big] \geq 0\big).$$

- No unique target function. No simple expression.

Empirical Risk Minimization

Empirical Risk Minimization

Idea: Replace the population distribution of the data by the empirical distribution of the training data. Given a training set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, we define the

Empirical Risk

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Empirical Risk Minimization principle

- consists in minimizing the empirical risk.

Problem: The target function for the empirical risk is only defined at the training points.

Linear regression

Linear regression

- We consider the OLS regression for the linear hypothesis space.
- We have $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \mathbb{R}$ and ℓ the square loss.

Consider the hypothesis space:

$$S = \{f_{\mathbf{w}} \mid \mathbf{w} \in \mathbb{R}^p\} \qquad \text{with} \qquad f_{\mathbf{w}} : \mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x}.$$

Given a training set $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ we have

$$\widehat{\mathcal{R}}_n(f_w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

with

- the vector of outputs $\mathbf{y}^\top = (y_1, \dots, y_n) \in \mathbb{R}^n$
- the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ whose ith row is equal to $\mathbf{x}_i^\top$.

Solving linear regression

To solve $\min_{\mathbf{w}\in\mathbb{R}^p} \widehat{\mathcal{R}}_n(f_{\mathbf{w}})$, we consider that

$$\widehat{\mathcal{R}}_n(f_w) = \frac{1}{2n}\left(\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\,\mathbf{w}^\top \mathbf{X}^\top \mathbf{y} + \|\mathbf{y}\|^2\right)$$

is a differentiable convex function whose minima are thus characterized by the

Normal equations

$$\boxed{\mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y} = \mathbf{0}}$$

If $\mathbf{X}^\top \mathbf{X}$ is invertible, then $\widehat{f}$ is given by:

$$\widehat{f} : \mathbf{x}' \mapsto \mathbf{x}'^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Problem: $\mathbf{X}^\top \mathbf{X}$ is never invertible for $p > n$ and thus the solution is not unique.

Classification and plug-in predictors

Classification and plug-in predictors

Input space $\mathcal{X}$, output space $\mathcal{Y} = \{-1, 1\}$.

- Empirical risk for 0-1 loss and $\gamma : \mathcal{X} \to \{-1, 1\}$

$$\widehat{\mathcal{R}}_n^{0\text{-}1}(\gamma) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{\gamma(x_i) \neq y_i\}}$$

$\to$ Relax empirical risk to allow for real valued predictors

- Empirical risk for 0-1 loss and $f : \mathcal{X} \to \mathbb{R}$

$$\widehat{\mathcal{R}}_n^{0\text{-}1}(f) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{y_i f(x_i) \leq 0\}}$$

- Then use the plug in rule $\gamma(x_i) = \text{sign}(f(x_i))$.

$\to$ Problem: ER is non-convex, discontinuous

$\to$ NP-hard to optimize...

Classification *via* OLS regression

For regression, but assuming $Y \in \{-1, 1\}$

- the risk is

$$\mathbb{E}\left[\left(f(X) - Y\right)^2\right] = \mathbb{E}\left[\left(1 - Yf(X)\right)^2\right]$$

- the target function is

$$\mathbb{E}[Y|X] = f^*(X) \quad \text{with} \quad f^*(X) = 2\,\mathbb{P}(Y = 1|X) - 1$$

- the excess risk is $\mathbb{E}\left[\left(f(X) - f^*(X)\right)^2\right]$

For classification

- the target function is

$$\arg\max_{y \in \{-1,1\}} \mathbb{P}(Y = y|x = x) = \text{sign}(f^*(x))$$

Plug-in principle

- Learn $\widehat{f}(x)$ using OLS regression
- Use the plug-in predictor for classification $\widehat{y} := \widehat{\gamma}(x) = \text{sign}(\widehat{f}(x))$

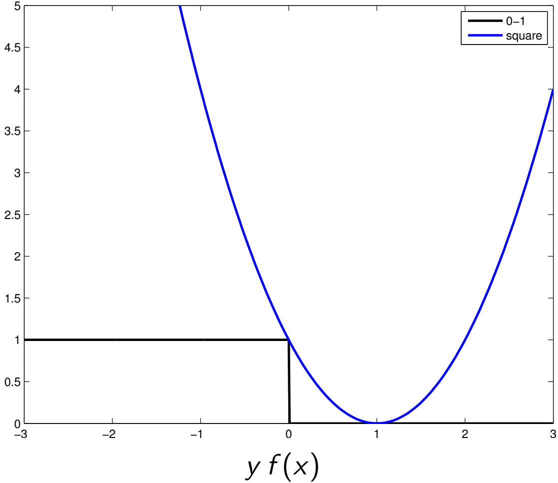
Zero one loss *vs* square loss

0-1 loss

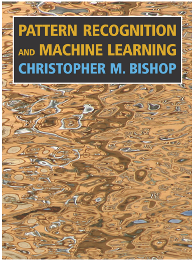
$$\ell(f(x), y) = 1_{\{y\,f(x) \leq 0\}}$$

Square loss

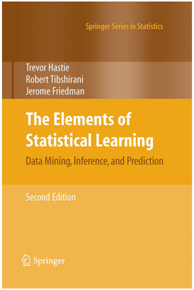
$$\ell(f(x), y) = (1 - y\,f(x))^2$$



References

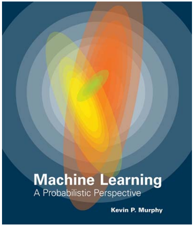


Pattern Recognition and Machine Learning,
Christopher Bishop, Springer (2006).
`http://research.microsoft.com/~cmbishop/PRML/`

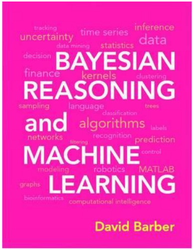


The Elements of Statistical Learning,
Trevor Hastie, Rob Tibshirani, Jerome Friedman,
Springer (2010).
`http://statweb.stanford.edu/~tibs/ElemStatLearn/`

References II



Machine Learning, a probabilistic perspective
Kevin Murphy, MIT Press (2012).



Bayesian reasoning and machine learning,
David Barber,
Cambridge University Press (2012).
`http://www.cs.ucl.ac.uk/staff/d.barber/brml/`