FOUNDATIONS OF STATISTICAL ANALYSIS & MACHINE LEARNING

Amric Trudel
amric.trudel@epita.fr

COURSE PROGRAM

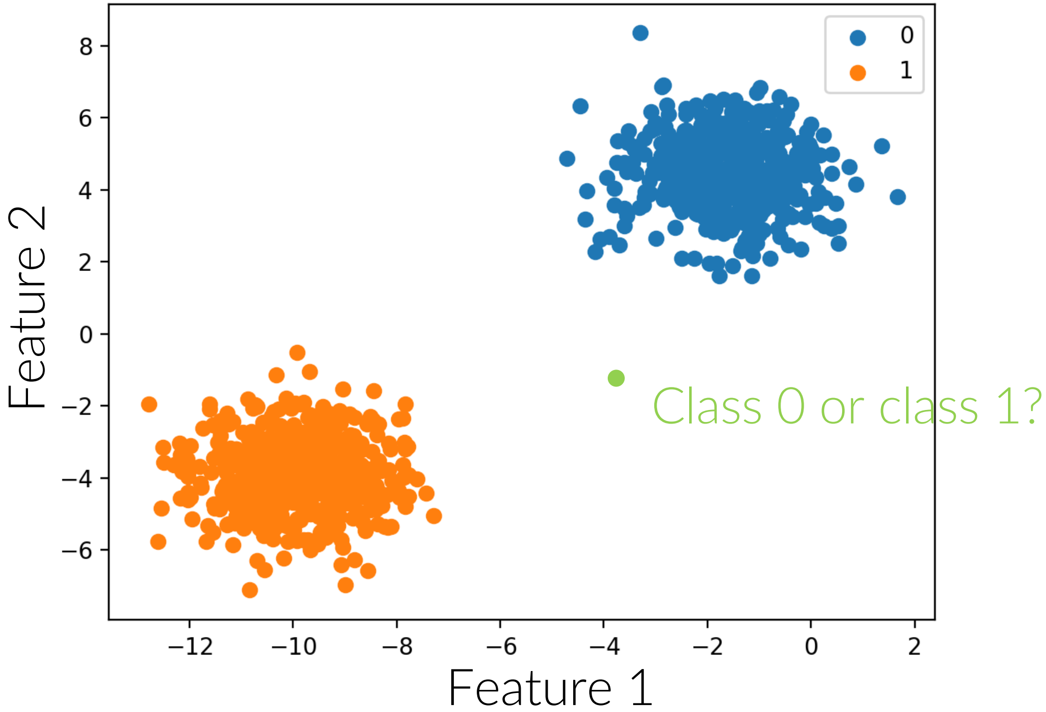
Structure

PREPARATION	Data exploration
	Data preprocessing
REGRESSION	Linear regression with one variable
	Multiple and polynomial regression
CLASSIFICATION	Logistic regression
	Classification model assessment
	k-NN, Decision Tree, SVM
CLUSTERING	k-means, hierarchical clustering
DIMENSIONALITY REDUCTION	Principal Components Analysis
ALL NOTIONS	Final assignment

CLASSIFICATION

Problem statement

Target = Categories (**Discrete values)**



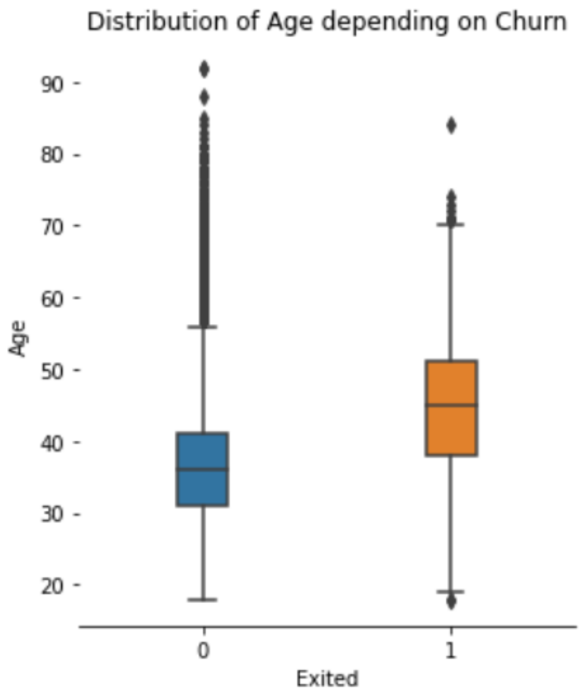
CLASSIFICATIONS

General approach

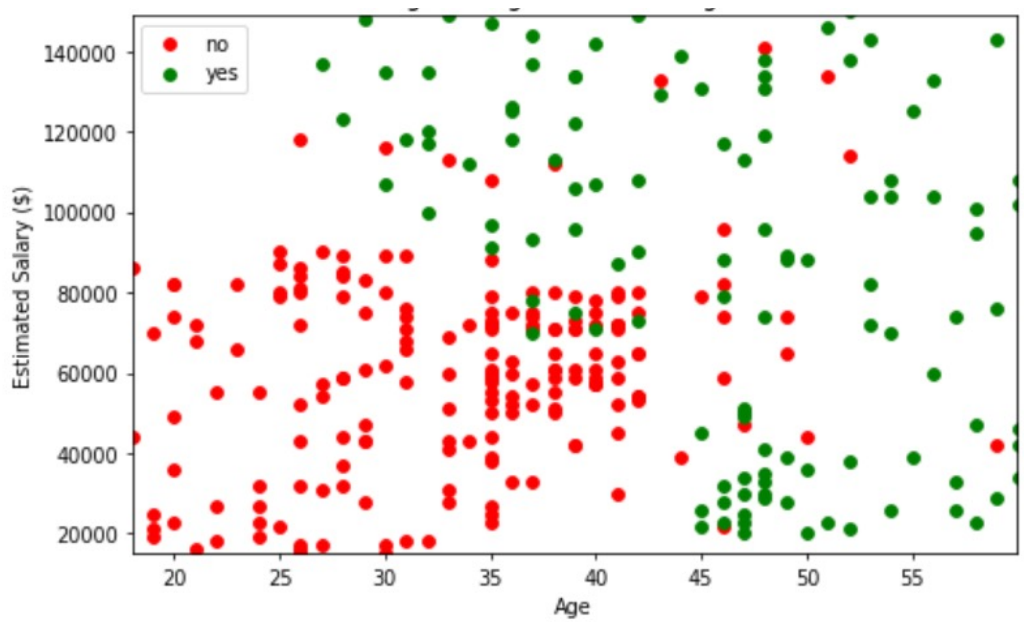
1. Get some intuition from **data inspection** (visualization, correlation, etc.)
2. **Choose a model**
3. **Find the parameters** that minimize a criteria (cost function)
4. **Evaluate the performance**

CLASSIFICATIONS

Data inspection



Distribution comparison



Distribution visualization

CLASSIFICATIONS

Model choice

- Logistic regression
- Discriminant analysis
- k-Nearest Neighbors
- Decision Tree
- Support Vector Machines
- …

- Data transformation

CLASSIFICATIONS

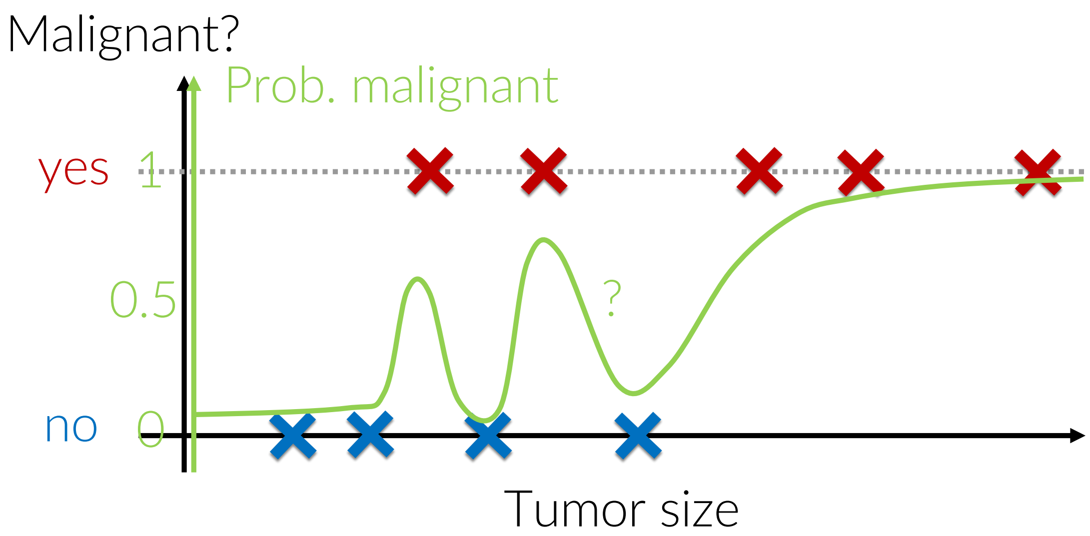
Performance evaluation

- Confusion matrix
- Accuracy, precision, recall
- ROC curve, AUC

LOGISTIC REGRESSION

Model definition

- Rather than modeling the response Y directly, the objective is to model the probability that Y belongs to a particular category.

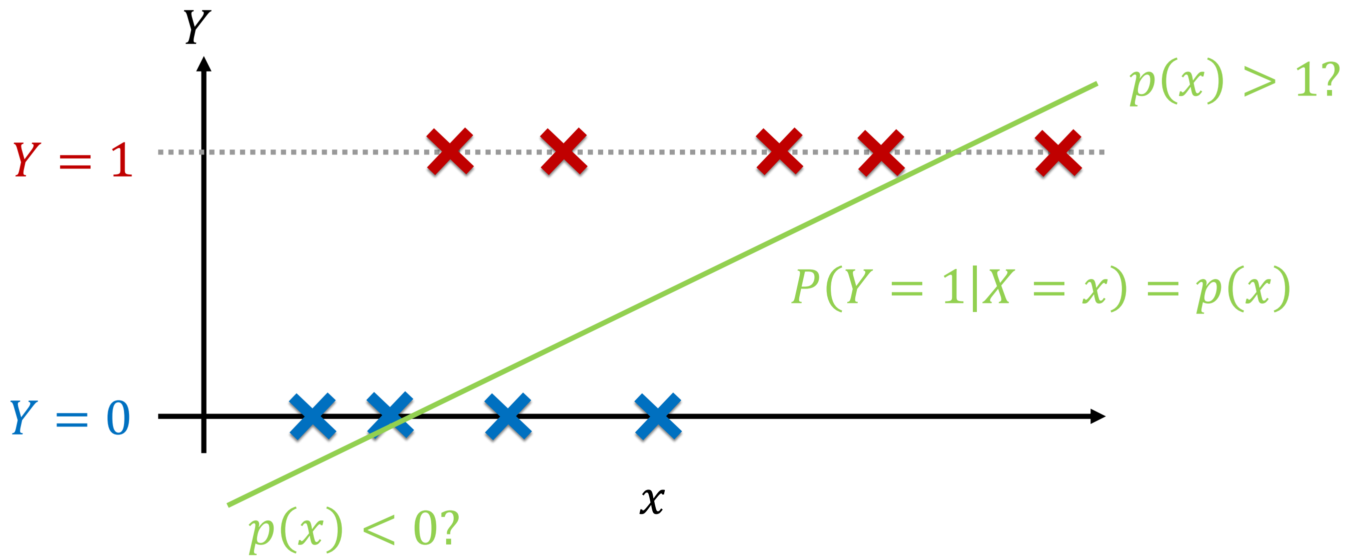


LOGISTIC REGRESSION

Model definition

- The linear regression model is not (directly) usable here:

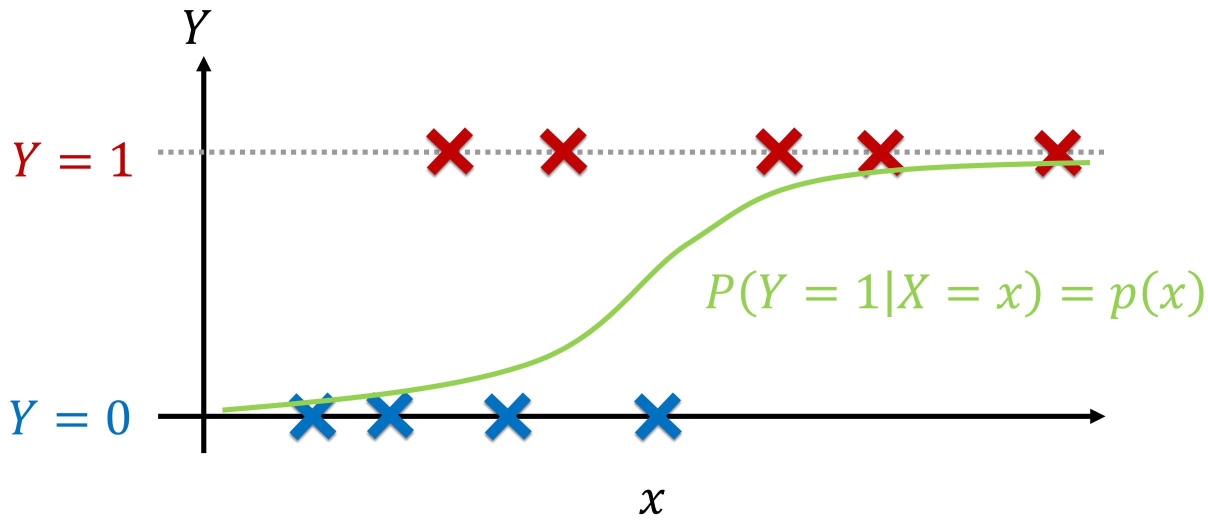
$$p(x) = \beta_0 + \beta_1 x$$



LOGISTIC REGRESSION

Model definition

- We encapsulate the function $z = \beta_0 + \beta_1 x$ so to map $[0,1]$.
- Logistic function: $p(x) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$



LOGISTIC REGRESSION

Model fitting

- **Maximum Likelihood Method**: $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen such that the predicted probability $\hat{p}(x_i)$ of the response for each individual corresponds as closely as possible to the individuals's observed response status.
- Sample elements are assumed to be independent Bernoulli variables:

$$Y = \begin{cases} 1, \text{with probability} & p \\ 0, \text{wih probability} & 1-p \end{cases}$$

or equivalently:

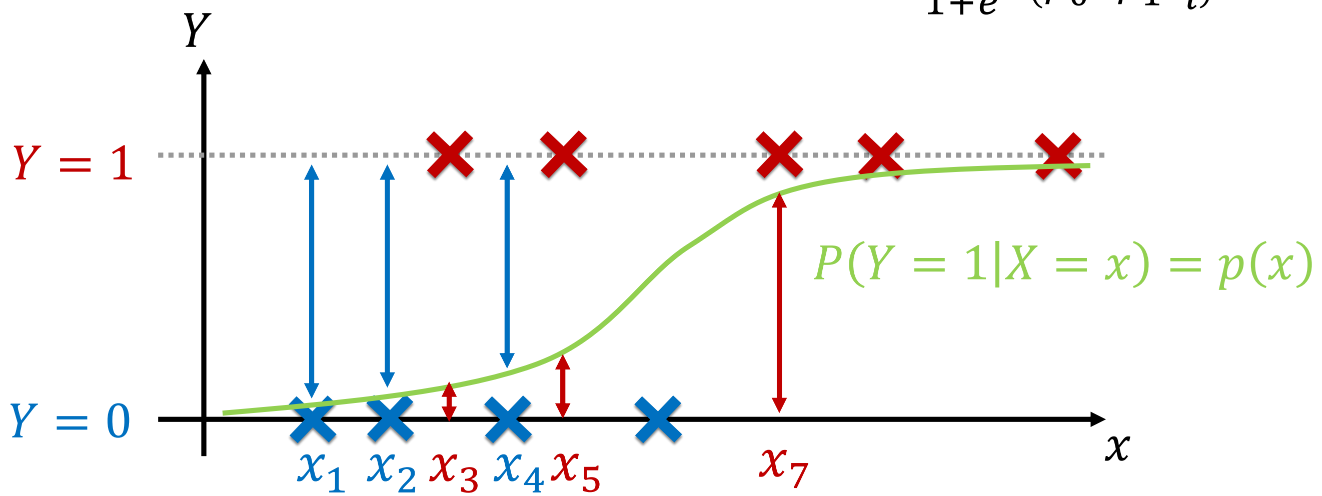
$$\mathrm{P}[Y = y] = p^{y}(1-p)^{1-y} \text{ with } y \in \{0,1\}$$

LOGISTIC REGRESSION

Model fitting

- **Likelihood function to maximize:** $L(\beta_0, \beta_1) = \prod_{i=1}^{m} p(x_i)^{y_i}\left(1 - p(x_i)\right)^{1-y_i}$

with $p(x_i) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$



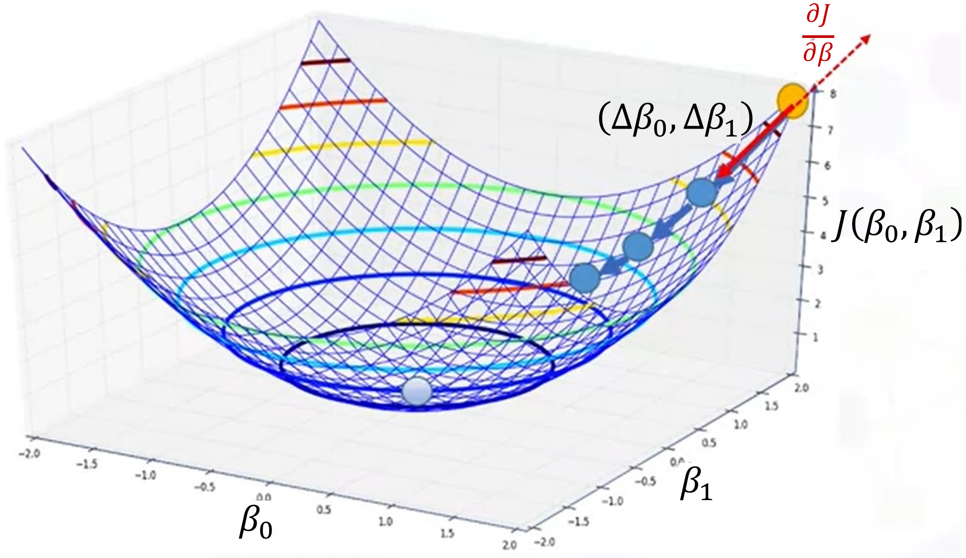
$$L(\beta_0, \beta_1) = \left(1 - p(x_1)\right)\times\left(1 - p(x_2)\right)\times p(x_3)\times\left(1 - p(x_4)\right)\times p(x_5)\times\cdots$$

LOGISTIC REGRESSION

Model fitting

- Cost function: $J(\beta_0, \beta_1) = -\frac{1}{m}\sum_{i=1}^{m} Y_i \log\big(p(x_i)\big) + (1 - Y_i)\log\big(1 - p(x_i)\big)$

- Optimization method: Gradient Descent
  - Arbitrary start $(\hat{\beta}_0, \hat{\beta}_1)$
  - Computation of the gradient at that position to determine the position (direction + range) of the next position
  - Iteration until a stopping criterion is reached



LOGISTIC REGRESSION

Python implementation

- **Feature scaling is important for classification!**

- Training the model:

```
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression()
classifier.fit(X_train, y_train)
```

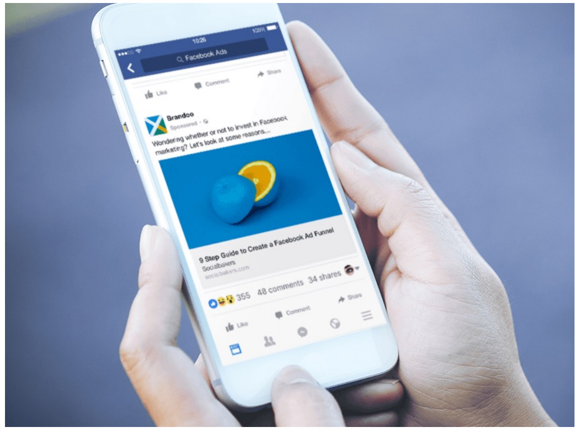
- Using the model for predicting:

```
y_proba = classifier.predict_proba(X_test)# probabilities
y_pred = classifier.predict(X_test)  # class predictions
```

LOGISTIC REGRESSION

Example of implementation

- Data set: user profiles and sales information
- Objectives:
  - Train a logistic regression to predict purchase based on profile information



	User ID	Gender	Age	EstimatedSalary	Purchased
1	15624510	Male	19	19000	no
2	15810944	Male	35	20000	no
3	15668575	Female	26	43000	no
4	15603246	Female	27	57000	no
5	15804002	Male	19	76000	no
6	15728773	Male	27	58000	no
7	15598044	Female	27	84000	no
8	15694829	Female	32	150000	yes
9	15600575	Male	25	33000	no
10	15727311	Female	35	65000	no
11	15570769	Female	26	80000	no
12	15606274	Female	26	52000	no
13	15746139	Male	20	86000	no

LOGISTIC REGRESSION

Student practice

- Data set: breast cancer diagnosis based on tumor characteristics
- Objectives:
  - Train logistic regressions
  - Visualize results (probability, predictions)

mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	tumor type
14.69	13.98	98.22	656.1	0.10310	0.18360	0.14500	0.06300	0.2086	0.07406	benign
13.17	18.66	85.98	534.6	0.11580	0.12310	0.12260	0.07340	0.2128	0.06777	malignant
12.95	16.02	83.14	513.7	0.10050	0.07943	0.06155	0.03370	0.1730	0.06470	benign
18.31	18.58	118.60	1041.0	0.08588	0.08468	0.08169	0.05814	0.1621	0.05425	malignant
15.13	29.81	96.71	719.5	0.08320	0.04605	0.04686	0.02739	0.1852	0.05294	malignant
16.16	21.54	106.20	809.8	0.10080	0.12840	0.10430	0.05613	0.2160	0.05891	malignant
19.19	15.94	126.30	1157.0	0.08694	0.11850	0.11930	0.09667	0.1741	0.05176	malignant
18.08	21.84	117.40	1024.0	0.07371	0.08642	0.11030	0.05778	0.1770	0.05340	malignant