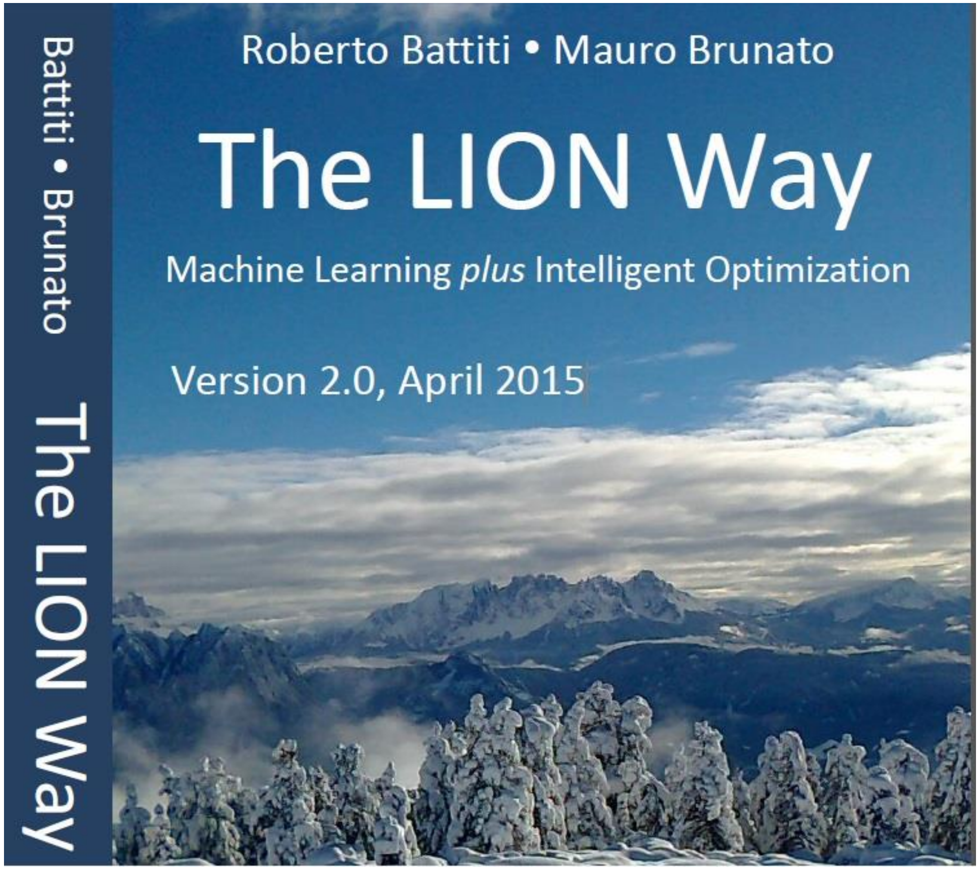


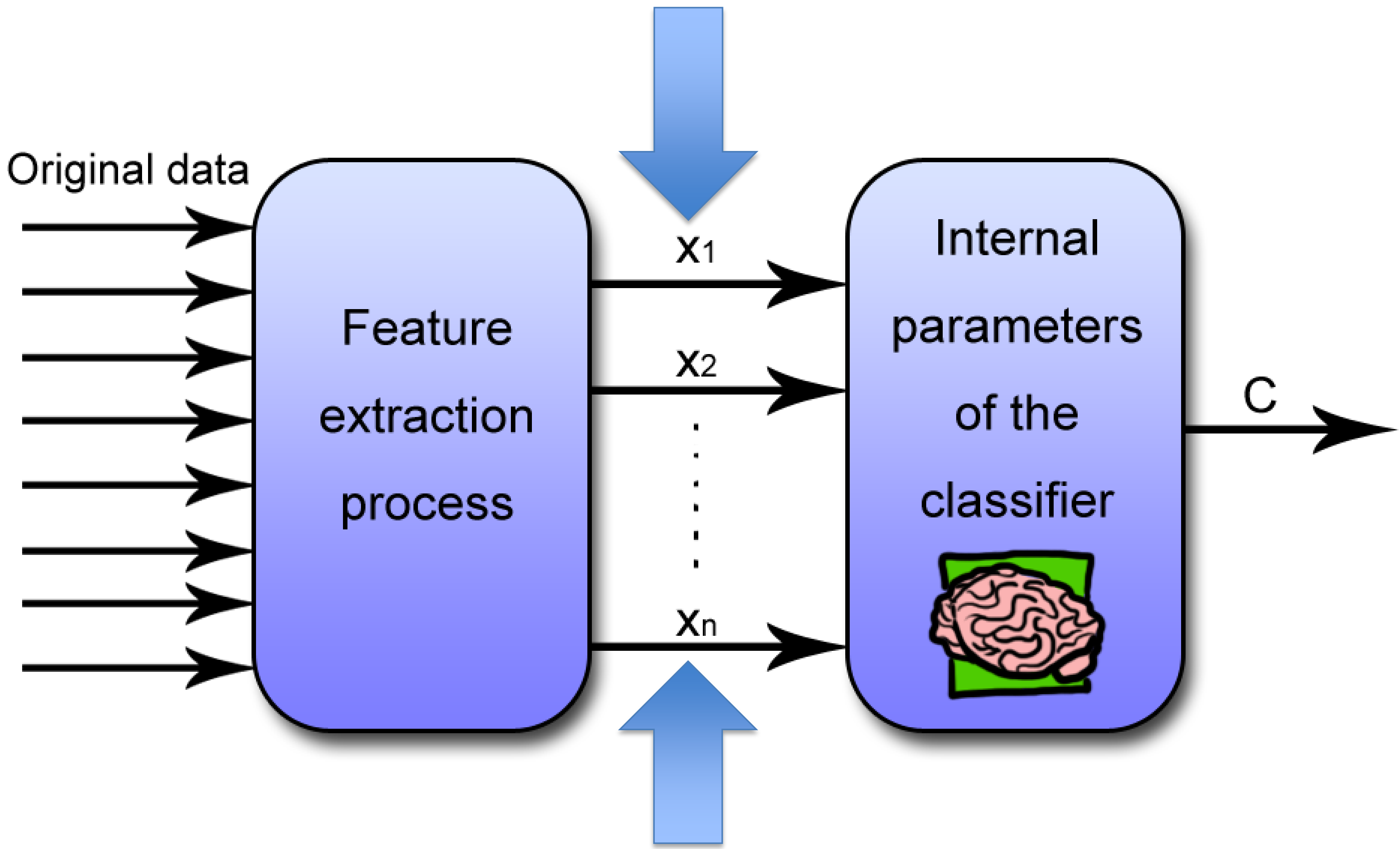
ROBERTO BATTITI, MAURO BRUNATO.
The LION Way: Machine Learning plus *Intelligent Optimization*.
LIONlab, University of Trento, Italy,
Apr 2015

http://intelligent-optimization.org/LIONbook

Chap.7 Ranking and selecting features

I don’t mind my eyebrows. They add. . . something to me. I wouldn’t say they were my best feature, though. People tell me they like my eyes. They distract from the eyebrows.
(Nicholas Hoult)

Feature selection



Feature selection (2)

- Before starting to learn a model from the examples, one must be sure that the input data have sufficient information to predict the outputs, without excessive redundancy, which may causes “big” models and poor generalization

- Feature selection is the process of selecting a subset of relevant features to be used in model construction.

Reasons for feature selection

- Selecting a small number of informative features has advantages:

1. Dimensionality reduction
2. Memory usage reduction
3. Improved generalization
4. Better human understanding

Methods for feature selection

- Feature selection is a problem with many possible solutions: no simple recipe.

1. Use the designer **intuition and existing knowledge**
2. **Estimate the relevance or discrimination power** of the individual features

Wrapper, Filter and Embedded methods

- The **value of a feature is related to a model-construction method**. Three classes of methods:

1. **Wrapper methods** are built "around" a specific predictive model (measure error rate)
2. **Filter methods** use a **proxy measure** instead of the error rate to score a feature subset
3. **Embedded methods** perform feature selection as an integral part of the model construction process.

Top-down and Bottom-up methods

- In a **bottom-up** method one gradually **adds** the ranked features in the order of their individual discrimination power and stops when the error rate stops decreasing
- In a **top-down truncation** method one starts with the complete set of features and progressively **eliminates** features while searching for the optimal performance point

Linear models

Can we associate the importance of a feature to its weight?

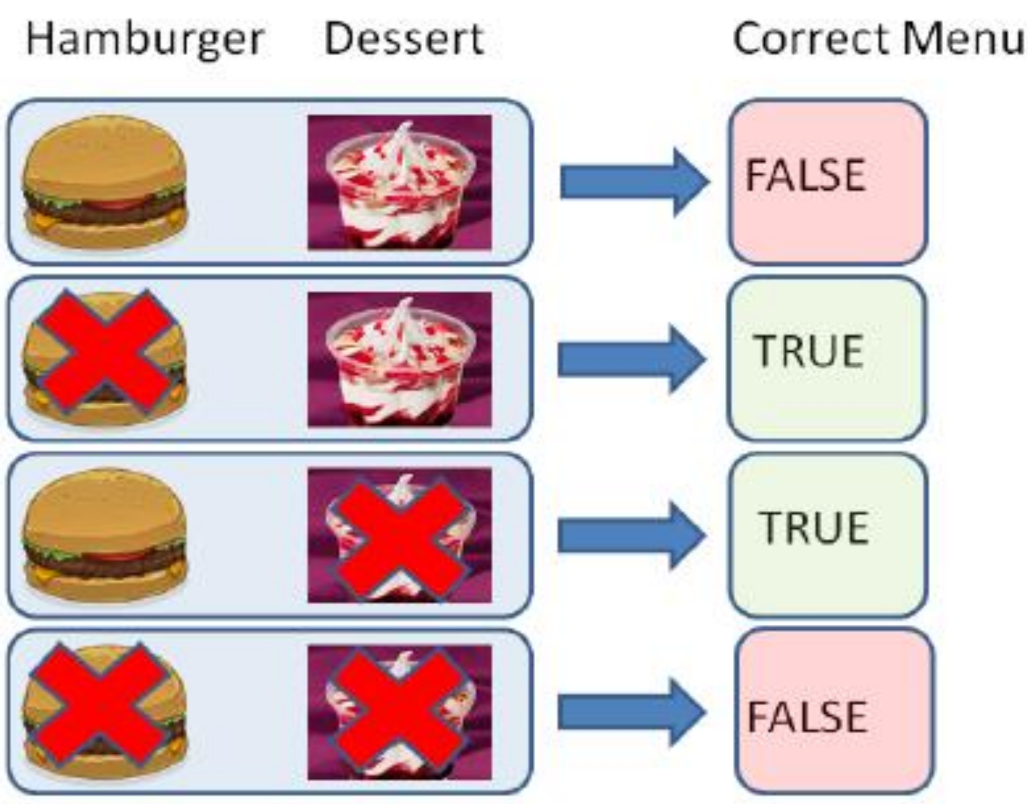
$$y = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d.$$

Careful with ranges and scaling.

Normalization helps.

Nonlinearities and mutual relationships between features

Measuring individual features in isolation will discard **mutual relationships** → selection can be suboptimal



XOR function of two inputs

E.g., to get a proper meal one needs to eat either a hamburger or a dessert but not both.

The individual presence or absence of a hamburger (or of a dessert) in a menu will not be related to classifying a menu as correct or not.

Correlation coefficient

Pearson correlation coefficient: widely used measure of linear relationship between numeric variables.

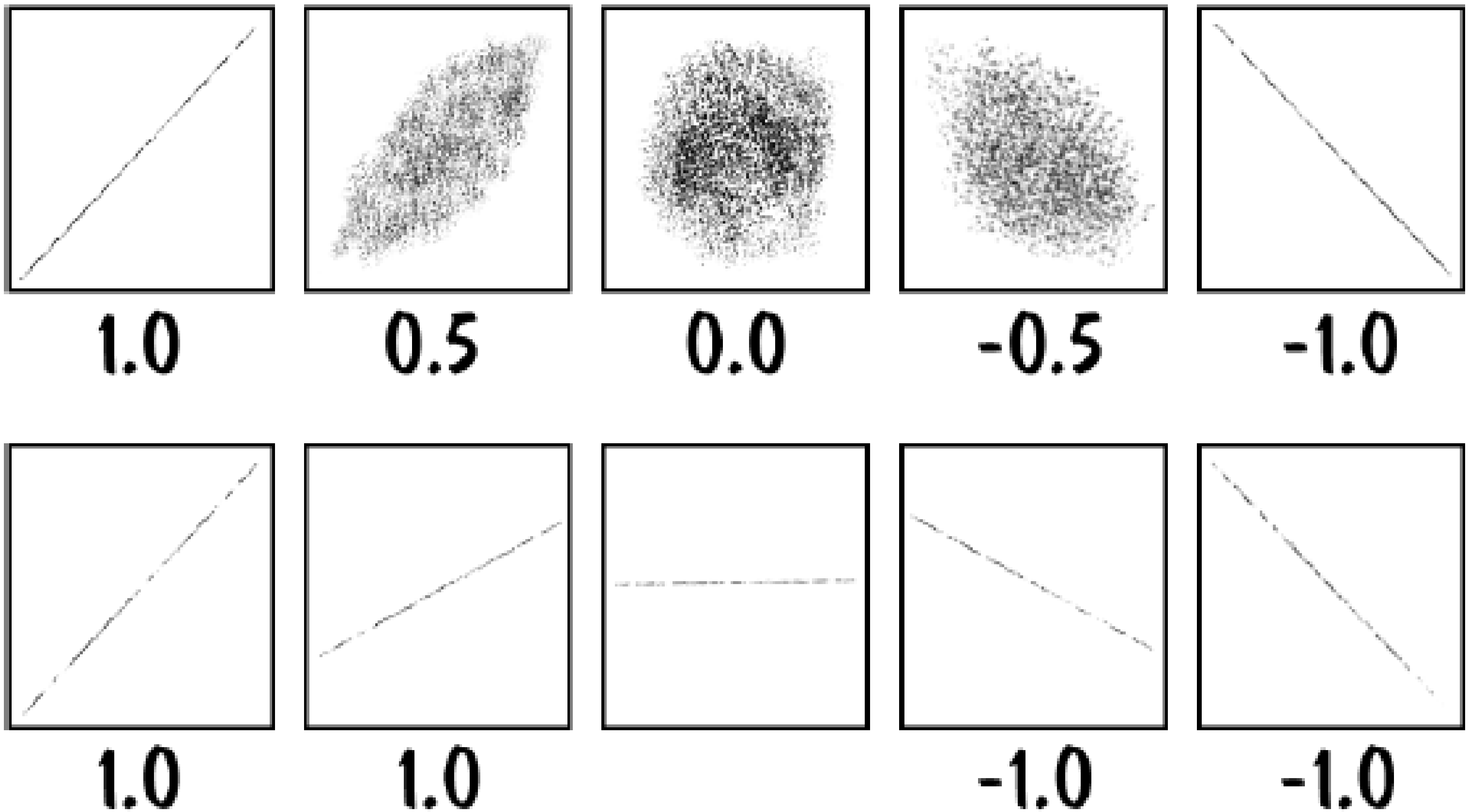
Y random variable associated with the output

Xi random variable associated with an input

$$\rho_{X_i,Y} = \frac{\mathrm{cov}[X_i, Y]}{\sigma_{X_i}\sigma_Y} = \frac{E[(X_i - \mu_{X_i})(Y - \mu_Y)]}{\sigma_{X_i}\sigma_Y};$$

Examples of data distributions and corresponding correlation values

Correlation coefficient (2)



Examples of data distributions and corresponding correlation values

Correlation Ratio

- **Correlation ratio** is used to measure a relationship between a numeric input and a **categorical** output.
- significant →at least one outcome class where the feature's average value is *significantly different* from the average on all classes
- Let L_y be the number of times that **outcome y** appears, so that one can **partition the sample input vectors** by their output:

$$\forall y \in Y \qquad S_y = \left((x_{jy}^{(1)}, \ldots, x_{jy}^{(n)}); j = 1, \ldots, \ell_y\right).$$

Inpuuts leading to output *y*

Correlation ratio (2)

- Average of the i-th feature *within* each output class:

$$\forall y \in Y \qquad \bar{x}_y^{(i)} = \frac{1}{\ell_y} \sum_{j=1}^{\ell_y} x_{jy}^{(i)},$$

- Overall average:

$$\bar{x}^{(i)} = \frac{1}{\ell} \sum_{y \in Y} \sum_{j=1}^{\ell_y} x_{jy}^{(i)} = \frac{1}{\ell} \sum_{y \in Y} \ell_y \bar{x}_y^{(i)}.$$

- Correlation ratio between the i-*th* feature and outcome:

$$\eta^2_{X_i,Y} = \frac{\sum_{y \in Y} \ell_y (\bar{x}_y^{(i)} - \bar{x}^{(i)})^2}{\sum_{y \in Y} \sum_{j=1}^{\ell_y} (x_{jy}^{(i)} - \bar{x}^{(i)})^2}.$$

Statistical hypothesis testing

- A statistical hypothesis test is a method of making statistical decisions by using experimental data.
- Hypothesis testing answers the question: Assuming that the **null hypothesis** is true, ***what is the probability of observing a value for the test statistic that is at least as large as the value that was actually observed***? Reject if prob. is too low.

- Statistically significant ←→ **unlikely to have occurred by chance.**

Relationship between two categorical features

- **Null hypothesis** that the two events “occurrence of term t” and “document of class c” are **independent**, the expected value of the above counts for joint events are obtained by **multiplying probabilities** of individual events
- If the count deviates from the one expected for two independent events, one can conclude that the two events are **dependent**, and that therefore the feature is significant to predict the output. Check if the deviation is sufficiently large that it cannot happen by chance.

Chi-squared test

- Chi-squared statistic:

If independent

$$\chi^2 = \sum_{c,t} \frac{[\text{count}_{c,t} - n \cdot \Pr(\text{class} = c) \cdot \Pr(\text{term} = t)]^2}{n \cdot \Pr(\text{class} = c) \cdot \Pr(\text{term} = t)}.$$

- where $\text{count}_{c,t}$ is the number of occurrences of the value t given the class c
- the best features are the ones with larger χ^2 values

Mutual information (1): Entropy

- The uncertainty in an output distribution can be measured from its entropy:

$$H(Y) = -\sum_{y \in Y} \Pr(y) \log \Pr(y).$$

- After knowing a specific input value x, the uncertainty in the output can decrease

Mutual information (2): Conditional Entropy

- The entropy of Y **after knowing the i-th input feature value** is

$$H(Y|x_i) = -\sum_{y \in Y} \Pr(y|x_i) \log \Pr(y|x_i),$$

- The **conditional entropy** of variable Y is the expected value of $H(Y|x_i)$

$$H(Y|X_i) = E_{x_i \in X_i}[H(Y|x_i)] = \sum_{x_i \in X_i} \Pr(x_i) H(Y|x_i).$$

Mutual Information (3)

- **Mutual information between X_i and Y** :

The amount by which the uncertainty **decreases**

$$I(X_i;Y) = I(Y;X_i) = H(Y) - H(Y|X_i).$$

- An equivalent expression which clarifies the symmetry between Xi and Y:

$$I(X_i;Y) = \sum_{y,x_i} \Pr(y,x_i) \log \frac{\Pr(y,x_i)}{\Pr(y)\Pr(x_i)}.$$

- Mutual Information captures **arbitrary non-linear dependencies** between variables

GIST

- Reducing the number of input attributes used by a model, while keeping roughly equivalent performance, has many advantages.

- It is difficult to rank individual features without considering the specific modeling method and their mutual relationships.

GIST 2

- Trust the correlation coefficient only if you have reasons to suspect linear relationships
- Correlation ratio can be computed even if the outcome is not quantitative
- Use chi-square to identify possible dependencies between inputs and output
- Use mutual information to estimate **arbitrary dependencies** between qualitative or quantitative features