# Machine learning

## Probabilistic learning

Corso di Laurea Magistrale in Informatica

Università di Roma Tor Vergata

Prof. Giorgio Gambosi

a.a. 2021-2022

Qui il risultato della predizione è $p(t|x)$.

As done before, we assume that the observed dataset (features and target) has been derived by randomly sampling:

- $\mathscr{X}$ according to the probability distribution $p_{\mathscr{D}_1}(x)$ (usually the uniform distribution)
- $\mathscr{Y}$ according to the conditional distribution $p_{\mathscr{D}_2}(y|x)$

1. we may then consider a class of possible conditional distributions $\mathscr{P}$ and
2. select (infer) the "best" conditional distribution $p^* \in \mathscr{P}$ from the available knowledge (that is, the dataset), according to some measure $q$
3. given any new item $x$, apply $p^*(y|\mathbf{x})$ to assign probabilities for each possible value of the corresponding target
4. an independent decision strategy must be applied to $p^*(y|\mathbf{x})$ to return a specific prediction $h(\mathbf{x})$

- Dobbiamo poi cercare una misura equivalente alla loss a partire dal dataset. Supponiamo di avere una feature x, un valore target t, vogliamo predire il valore del target data la feature. Possiamo considerare l'insieme delle funzioni Gaussiane univariate, che differiranno fra loro sulla base di un parametro, ovvero la media. A questo punto possiamo dire che ci aspettiamo che, fissato un x, la distribuzione dei valori di target collegati ad un elemento che ha quel valore della feature abbia una distribuzione Gaussiana.

- Consideriamo la regressione lineare: vogliamo, dato un vettore x delle feature, predire un valore t del target. Ne consideriamo una dove abbiamo una sola feature e supponiamo di voler determinare il valore di t a partire dal valore di x usando un modello lineare. Quindi il valore predetto $y = w_1 \cdot x + w_0$ , con l'approccio funzionale, tutte le h sarebbero fatte in quel modo ed io sceglierei la migliore.
Con un approccio probabilistico vogliamo avere una distribuzione di probabilità per i valori del target: definiamo quindi una classe di distribuzioni, ad esempio Gaussiane, supponiamo che abbiano tutte la stessa varianza e cambi solo la media. Queste distribuzioni sono tutte della stessa forma ma con media diversa, quindi la classe delle nostre distribuzioni può essere quella delle distribuzioni Gaussiane che saranno parametriche nella media, ma devono essere dipendenti da x (in quanto distribuzioni condizionate da x), quindi possiamo dire che $p(y|x)$ è Gaussiana, di media che dipende da x ed è ad esempio proprio $w_1 \cdot x + w_0$ e varianza data. Quindi, fissato x, nel caso di funzioni avevamo una retta e di tutte le possibili rette cercavamo quella che si comportava meglio da un punto di vista della predizione. Nel caso probabilistico, fissato x, fissando $w_0$ e $w_1$ ci ritroviamo una distribuzione Gaussiana centrata su $w_1 \cdot x + w_0$ e che ha un'ampiezza predefinita. Quindi t si trova distribuito su una Gaussiana centrata nel valore y e non è esattamente il punto y, come nel caso precedente, abbiamo quindi dei valori di probabilità.
A questo punto, nell'approccio funzionale abbiamo un valore predetto ed uno corretto e quindi l'errore è tipicamente la distanza fra questi due punti. In questo ambito, per stimare la distanza possiamo dire che tanto minore è la distanza fra due elementi e tanto maggiore è il valore di probabilità. In questa distribuzione, se t fosse esattamente pari al valore medio, allora la probabilità sarebbe alta, mentre se fosse lontano sarebbe bassa.

- Vogliamo quindi trovare la probabilità di t: $N(t|w_0 + w_1 \cdot x, \sigma^2)$ quindi date media e varianza, qual è la probabilità del target. In questo modo la probabilità del target è inversamente proporzionale alla distanza. Quindi, la migliore distribuzione, se abbiamo tante distribuzioni ed un dataset è quella per il quale la probabilità su quel dataset è la più alta possibile.

# Inferring a best distribution

- how to define the class of possible conditional distributions $p(y|\mathbf{x})$?
  - usually, parametric approach: distributions defined by a common (arbitrary) structure and a set of parameters
- what is a measure $q(p, \mathcal{T})$ of the quality of the distribution (given the dataset $\mathcal{T} = (\mathbf{X}, \mathbf{t})$)?
  - this is related to how a dataset generated by randomly sampling from $\mathcal{D}_1$ (usually uniform) and $\mathcal{D}_2$ could be similar to the available dataset $\mathcal{T}$

# A different approach

Instead of finding a best distribution $p^* \in \mathscr{P}$ and use it to predict target probabilities as $p^*(y|\mathbf{x})$ for any element $\mathbf{x}$, we could

- consider for each possible conditional distribution $p \in \mathscr{P}$ its quality $q(p, \mathscr{T})$
- compose all conditional distributions $p(y|\mathbf{x})$ each weighted by its quality $q(p, \mathscr{T})$ (for example by means of a weighted averaging)
- apply the resulting distribution

# Different strategies

Assume $q$ takes the form of a probability distribution (of probability distribution)

- first approach: take the modal value (the distribution of maximum quality) and apply it to perform predictions
- second approach: compute the expectation of the distributions, wrt the probability distribution $q$

# Inference of predictive distribution

## Dataset

We assume elements in $\mathcal{T}$ correspond to a set of $n$ samples, independently drawn from the same probability distribution (that is, they are independent and identically distributed, i.i.d): they can be seen as $n$ realizations of a single random variable.

We are interested in learning, starting from $\mathcal{T}$, a predictive distribution $p(\mathbf{x}|\mathbf{X})$ (or $p(\mathbf{x}, t|\mathbf{X}, \mathbf{t})$) for any new element (or element-target pair). We may interpret this as the probability that, in a random sampling, the element actually returned is indeed $\mathbf{x}$ (or $\mathbf{x}, t$).

- in the case that $\mathcal{T} = \mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, we are interested in deriving the probability distribution $p(\mathbf{x}|\mathbf{X})$ of a new element, given the knowledge of the set $\mathbf{X}$
- in the case that $\mathcal{T} = (\mathbf{X}, \mathbf{t}) = \{(\mathbf{x}_1, t_1), \dots, (\mathbf{x}_n, t_n)\}$, we are interested in deriving the joint probability distribution $p(\mathbf{x}, t|\mathbf{X}, \mathbf{t})$ or, assuming $p(\mathbf{x}|\mathbf{X}, \mathbf{t})$ uniform and thus also independent from $\mathbf{X}, \mathbf{t}$, the conditional distribution $p(t|\mathbf{x}, \mathbf{X}, \mathbf{t})$, given the knowledge of the set of pairs $\mathbf{X}, \mathbf{t}$

# Probabilistic models

A probabilistic model is a collection of probability distributions with the same structure, defined over the data domain. Probability distribution are instances of the probabilistic model and are characterized by the values assumed by a set of parameters.

## Example

In a bivariate gaussian probabilistic model, distributions are characterized by the values assumed by:

1. the mean $\boldsymbol{\mu} = (\mu_1, \mu_2)$
2. the covariance matrix $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$

where $\sigma_{12} = \sigma_{21}$

A probabilistic model could be

**Parametric** if the set of parameters is given, finite, and independent from the data

**Non parametric** if the set of parameters is not given in advance, but derives from the data

# Bayesian learning at the model level

- Given a model space $\mathscr{M}$, let $m \in \mathscr{M}$ be a probabilistic model with parameters $\boldsymbol{\theta}$ ranging on a parameter space $\boldsymbol{\Theta}$.
- then, $p(\mathbf{x}|\boldsymbol{\theta}, m)$ is the predictive distribution from probabilistic model $m$ instantiated on parameter values $\boldsymbol{\theta}$
- Assume a prior parameter distribution $p(\boldsymbol{\theta}|m)$ is defined for the model.
- The corresponding prior predictive distribution is then

$$p(\mathbf{x}|m) = \int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, m) p(\boldsymbol{\theta}|m) d\boldsymbol{\theta}$$

## Bayesian learning at the model level

- Bayes' formula makes it possible to infer the posterior distribution of parameters, given the dataset $\mathcal{T}$

$$p(\boldsymbol{\theta}|\mathcal{T}, m) = \frac{p(\boldsymbol{\theta}|m)p(\mathcal{T}|\boldsymbol{\theta}, m)}{p(\mathcal{T}|m)} = \frac{p(\boldsymbol{\theta}|m)p(\mathcal{T}|\boldsymbol{\theta}, m)}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}'|m)p(\mathcal{T}|\boldsymbol{\theta}', m)d\boldsymbol{\theta}'}$$

- The posterior predictive distribution, given the model, is then

$$p(\mathbf{x}|\mathcal{T}, m) = \int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, m)p(\boldsymbol{\theta}|\mathcal{T}, m)d\boldsymbol{\theta}$$

This is usually very hard, if not impossible, to be done efficiently: two high-dimensional integrations to deal with.

- no analytical solutions, in general
- numerical solutions can be computationally expensive
- approximate solutions when possible

# Bayesian learning at the model level

- $p(\mathbf{x}|\boldsymbol{\theta}, m)$ is a specific predictive distribution in the collection defined by model $m$
- $p(\boldsymbol{\theta}|\mathcal{T}, m)$ is the probability of its parameter values given the observed dataset, it can be seen as a quality measure $q$ of the distribution wrt $\mathcal{T}$
- the predictive probability of an element $\mathbf{x}$ corresponds to the average of the distributions $p(\mathbf{x}|\boldsymbol{\theta}, m)$, weighted by the quality measure $p(\boldsymbol{\theta}|\mathcal{T}, m)$

# Bayesian learning of predictive distribution

Let $p(m)$ be any prior distribution of probabilistic models on model space $\mathscr{M}$

$$\sum_{m \in \mathscr{M}} p(m) = 1$$

In a bayesian framework, we may consider the posterior probability of each model

$$p(m|\mathscr{T}) = \frac{p(\mathscr{T}|m)p(m)}{p(\mathscr{T})}$$

The analytical expression of the predictive distribution turns out to be quite complex

$$
\begin{aligned}
p(\mathbf{x}|\mathcal{T}) &= \sum_{m \in \mathcal{M}} p(m|\mathcal{T}) p(\mathbf{x}|\mathcal{T}, m) = \sum_{m \in \mathcal{M}} p(m|\mathcal{T}) \int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, m) p(\boldsymbol{\theta}|\mathcal{T}, m) d\boldsymbol{\theta} \\
&= \sum_{m \in \mathcal{M}} p(m|\mathcal{T}) \int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, m) \frac{p(\boldsymbol{\theta}|m) p(\mathcal{T}|\boldsymbol{\theta}, m)}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}'|m) p(\mathcal{T}|\boldsymbol{\theta}', m) d\boldsymbol{\theta}'} d\boldsymbol{\theta} \\
&= \sum_{m \in \mathcal{M}} \frac{p(\mathcal{T}|m) p(m)}{p(\mathcal{T})} \int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, m) \frac{p(\boldsymbol{\theta}|m) p(\mathcal{T}|\boldsymbol{\theta}, m)}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}'|m) p(\mathcal{T}|\boldsymbol{\theta}', m) d\boldsymbol{\theta}'} d\boldsymbol{\theta} \\
&= \sum_{m \in \mathcal{M}} \frac{p(m)}{p(\mathcal{T})} \int_{\boldsymbol{\Theta}} p(\mathcal{T}|\boldsymbol{\theta}, m) p(\boldsymbol{\theta}|m) d\boldsymbol{\theta} \cdot \int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, m) \frac{p(\boldsymbol{\theta}|m) p(\mathcal{T}|\boldsymbol{\theta}, m)}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}'|m) p(\mathcal{T}|\boldsymbol{\theta}', m) d\boldsymbol{\theta}'} d\boldsymbol{\theta}
\end{aligned}
$$

Evaluating this expression seems unfeasible: how to make things simpler?

1. apply model inference

# Model inference

Model inference is the task of deriving, given a dataset $\mathcal{T}$ the "best" probability distribution defined on the same data domain, according to some quality measure

## Two phases

**Model selection** From a collection of possible probabilistic models select the probabilistic model $M$ best suited for $\mathcal{T}$

**Estimation** Given a probabilistic model $m$ with parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)$ derive the probability distribution (that is the assignment of values to $\boldsymbol{\theta}$) best suited for $\mathcal{T}$

# Model comparison

Instead of composing the predictions of all probabilistic models, select and apply the one which best suit wrt $\mathcal{T}$.

How to compare models? Use the posterior probability of each model, given the dataset

$$p(m|\mathcal{T}) = \frac{p(\mathcal{T}|m)p(m)}{p(\mathcal{T})}$$

Observe that:

- If we assume that no specific knowledge on probabilistic models is initially available, then the prior distribution is uniform.
- The evidence $p(\mathcal{T})$ is a constant with respect to $m$

As a consequence, $p(m|\mathcal{T}) \propto p(\mathcal{T}|m)$ and we may refer to the likelihood $p(\mathcal{T}|m)$ in order to compare models

## Validation

**Test set** Dataset is split into Training set (used for learning parameters) and Test set (used for measuring effectiveness). Good for large datasets: otherwise, small resulting training and test set (few data for fitting and validation)

**Cross validation** Dataset partitioned into $K$ equal-sized sets. Iteratively, in $K$ phases, use one set as test set and the union of the other $K-1$ ones as training set ($K$-fold cross validation). Average validation measures.

As a particular case, iteratively leave one element out and use all other points as training set (Leave-one-out cross validation).

Time consuming for large datasets and for models which are costly to fit.

## Information measures

Faster methods to compare model effectiveness, based on computing measures which take into account data fitting and model complexity.

Akaike Information Criterion (AIC) Let $\boldsymbol{\theta}$ be the set of parameters of the model and let $\boldsymbol{\theta}_{ML}$ be their maximum likelihood estimate on the dataset $\mathbf{X}$. Then,

$$AIC = 2|\boldsymbol{\theta}| - 2\log p(\mathbf{X}|\boldsymbol{\theta}_{ML}) = 2|\boldsymbol{\theta}| - 2\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{X})$$

lower values correspond to models to be preferred.

Bayesian Information Criterion (BIC) A variant of the above, defined as

$$\begin{aligned} BIC &= |\boldsymbol{\theta}| - \log|\mathbf{X}|2\log p(\mathbf{X}|\boldsymbol{\theta}_{ML}) \\ &= |\boldsymbol{\theta}|\log|\mathbf{X}| - 2\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{X}) \end{aligned}$$

Given a probabilistic model $m^*$, selected according to some approach, the predictive distribution turns out to be quite complex

$$p(\mathbf{x}|\mathcal{T}) \approx p(\mathbf{x}|\mathcal{T}, m^*) = \int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, m^*) p(\boldsymbol{\theta}|\mathcal{T}, m^*) d\boldsymbol{\theta}$$

$$= \int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, m^*) \frac{p(\boldsymbol{\theta}|m^*) p(\mathcal{T}|\boldsymbol{\theta}, m^*)}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}'|m^*) p(\mathcal{T}|\boldsymbol{\theta}', m^*) d\boldsymbol{\theta}'} d\boldsymbol{\theta}$$

- As noticed above, computing $p(\boldsymbol{\theta}|\mathcal{T}, m^*)$ and, from it, $p(\mathbf{x}|\mathcal{T}, m^*)$ can be quite hard if not impossible
- This leads to the idea of only estimating model inference that is the task of deriving, given $\mathcal{T}$ and $m^*$, the "best" probability distribution defined on the same data domain, according to some quality measure
- Only an estimate of the "best" value $\boldsymbol{\theta}^*$ in $\boldsymbol{\Theta}$ (according to some measure) is performed.
- The posterior predictive distribution can then be approximated as follows

$$p(\mathbf{x}|\mathcal{T}) \approx p(\mathbf{x}|\mathcal{T}, m^*) = \int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}, m^*) p(\boldsymbol{\theta}|\mathcal{T}, m^*) d\boldsymbol{\theta} \approx \int_{\boldsymbol{\Theta}} p(\mathbf{x}|\boldsymbol{\theta}^*, m^*) p(\boldsymbol{\theta}|\mathcal{T}, m^*) d\boldsymbol{\theta}$$
$$= p(\mathbf{x}|\boldsymbol{\theta}^*, m^*) \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}|\mathcal{T}) d\boldsymbol{\theta} = p(\mathbf{x}|\boldsymbol{\theta}^*, m^*)$$

# Likelihood

Given a dataset $\mathcal{T}$ and a probability distribution $p$ of parameters $\boldsymbol{\theta}$ defined on the same data domain,

- the likelihood of $\boldsymbol{\theta}$ wrt $\mathcal{T}$ is defined as
$$L(\boldsymbol{\theta}|\mathcal{T}) = p(\mathcal{T}|\boldsymbol{\theta})$$
the probability of the dataset (that the dataset is generated) under distribution $p$ with parameters $\boldsymbol{\theta}$
- while the probability $p(\mathcal{T}|\boldsymbol{\theta})$ is considered as a function of $p(\mathcal{T}|\boldsymbol{\theta})$ with $\boldsymbol{\theta}$ fixed, the likelihood $L(\boldsymbol{\theta}|\mathcal{T})$ is a function of $\boldsymbol{\theta}$ with $\mathcal{T}$ fixed
- parameters $\boldsymbol{\theta}$ are considered as (independent) variables (frequentist interpretation of probability)

- By assuming that elements in $\mathcal{T}$ are i.i.d.,

$$L(\boldsymbol{\theta}|\mathcal{T}) = p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta}) \qquad \text{in the first case}$$

$$L(\boldsymbol{\theta}|\mathcal{T}) = p(\mathbf{X}, \mathbf{t}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i, t_i|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(t_i|\mathbf{x}_i, \boldsymbol{\theta}) p(\mathbf{x}_i|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta}) \prod_{i=1}^{n} p(t_i|\mathbf{x}_i, \boldsymbol{\theta})$$

$$= p(\mathbf{x}) \prod_{i=1}^{n} p(t_i|\mathbf{x}_i, \boldsymbol{\theta}) \propto \prod_{i=1}^{n} p(t_i|\mathbf{x}_i, \boldsymbol{\theta}) \qquad \text{in the second case, assuming } p(\mathbf{x}|\boldsymbol{\theta}) \text{ uniform}$$

## Approach

Frequentist point of view: parameters are deterministic variables, whose value is unknown and must be estimated.
Determine the parameter value that maximize the likelihood

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}|\mathcal{T}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathbf{X}|\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})$$

or

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}|\mathcal{T}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathbf{X}, \mathbf{t}|\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathbf{x}) \prod_{i=1}^{n} p(t_i|\mathbf{x}_i, \boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{n} p(t_i|\mathbf{x}_i, \boldsymbol{\theta})$$

Not case Bayesians: $\underset{\theta}{\operatorname{argmax}}\; p(\theta|X)$

## Log-likelihood

$$l(\boldsymbol{\theta}|\mathcal{T}) = \ln L(\boldsymbol{\theta}|\mathcal{T})$$

is usually preferrable, since products are turned into sums, while $\boldsymbol{\theta}^*$ remains the same (since log is a monotonic function), that is

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, l(\boldsymbol{\theta}|\mathcal{T}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}|\mathcal{T})$$

## Estimate

$$\boldsymbol{\theta}^*_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathbf{X}|\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\boldsymbol{\theta})$$

or

$$\boldsymbol{\theta}^*_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathbf{X}, \mathbf{t}|\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{n} \ln p(t_i|\mathbf{x}_i, \boldsymbol{\theta})$$

## Solution

Solve the system

$$\frac{\partial l(\boldsymbol{\theta}|\mathcal{T})}{\partial \theta_i} = 0 \qquad\qquad i = 1, \dots, d$$

more concisely,

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathcal{T}) = \mathbf{0}$$

## Prediction

Probability of a new observation $\mathbf{x}$:

$$p(\mathbf{x}|\mathbf{X}) = \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta} \approx \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}^*_{ML}) p(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta} = p(\mathbf{x}|\boldsymbol{\theta}^*_{ML}) \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta} = p(\mathbf{x}|\boldsymbol{\theta}^*_{ML})$$

Predictive distribution $t|\mathbf{x}$:

$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}) = \int_{\boldsymbol{\theta}} p(t|\mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{t}) d\boldsymbol{\theta} \approx \int_{\boldsymbol{\theta}} p(t|\mathbf{x}, \boldsymbol{\theta}^*_{ML}) p(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta} = p(\mathbf{x}|\boldsymbol{\theta}^*_{ML}) \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{t}) d\boldsymbol{\theta} = p(t|\mathbf{x}, \boldsymbol{\theta}^*_{ML})$$

## Example

Collection $\mathbf{X}$ of $n$ binary events, modeled through a Bernoulli distribution with unknown parameter $\phi$

$$p(x|\phi) = \phi^x(1-\phi)^{1-x}$$

Likelihood: $L(\phi|\mathbf{X}) = \prod_{i=1}^{n} \phi^{x_i}(1-\phi)^{1-x_i}$

Log-likelihood: $l(\phi|\mathbf{X}) = \sum_{i=1}^{n} (x_i \ln\phi + (1-x_i)\ln(1-\phi)) = n_1 \ln\phi + n_0 \ln(1-\phi)$

where $n_0$ ($n_1$) is the number of events $x \in \mathbf{X}$ equal to 0 (1)

$$\frac{\partial l(\phi|\mathbf{X})}{\partial\phi} = \frac{n_1}{\phi} - \frac{n_0}{1-\phi} = 0 \qquad \Longrightarrow \qquad \phi^*_{ML} = \frac{n_1}{n_0+n_1} = \frac{n_1}{n}$$

## Example

Linear regression: collection $\mathbf{X}, \mathbf{t}$ of value-target pairs, modeled as $p(\mathbf{x}, t) = p(\mathbf{x})p(t|\mathbf{x}, \mathbf{w}, \sigma^2)$, with $\mathbf{w} \in \mathbb{R}^d$, $w_0 \in \mathbb{R}$:

- $p(\mathbf{x})$ uniform
- $p(t|\mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{w}^T\mathbf{x} + w_0, 1/\beta)$ ($\beta$, the inverse of the variance, is the precision)

Likelihood: $L(\mathbf{t}|\mathbf{X}, \mathbf{w}, w_0, \beta) = \prod_{i=1}^n p(t_i|\mathbf{x}_i, \mathbf{w}, w_0, \beta) = \prod_{i=1}^n \mathcal{N}(\mathbf{w}^T\mathbf{x}_i + w_0, \beta)$

Log-likelihood:

$$l(\mathbf{t}|\mathbf{X}, \mathbf{w}, w_0, \beta) = \sum_{i=1}^n \ln p(t_i|\mathbf{x}_i, \mathbf{w}, w_0, \beta) = \sum_{i=1}^n \ln\left(\sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta(\mathbf{w}^T\mathbf{x}_i + w_0 - t_i)^2}{2}}\right) = \sum_{i=1}^n \left(-\frac{\beta(\mathbf{w}^T\mathbf{x}_i + w_0 - t_i)^2}{2} + \frac{1}{2}\ln\beta - \frac{1}{2}\ln(2\pi)\right)$$

$$= -\frac{\beta}{2}\sum_{i=1}^n (\mathbf{w}^T\mathbf{x}_i + w_0 - t_i)^2 + \frac{n}{2}\ln\beta - \frac{n}{2}\ln(2\pi)$$

## Example

$$\frac{\partial}{\partial w_k} l(\mathbf{t}|\mathbf{X}, \mathbf{w}, w_0, \beta) = -\frac{\beta}{2} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i + w_0 - t_i) x_{ik} \qquad k = 1, \dots, d$$

$$\frac{\partial}{\partial w_0} l(\mathbf{t}|\mathbf{X}, \mathbf{w}, w_0, \beta) = -\frac{\beta}{2} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i + w_0 - t_i)$$

$$\frac{\partial}{\partial \beta} l(\mathbf{t}|\mathbf{X}, \mathbf{w}, w_0, \beta) = -\frac{1}{2} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i + w_0 - t_i)^2 + \frac{n}{2\beta}$$

The ML estimation for $\mathbf{w}, w_0$ (linear regression coefficients) is obtained as the solution of the $(d+1, d+1)$ linear system

$$\sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i + w_0 - t_i) x_{ik} = 0 \qquad k = 1, \dots, d$$

$$\sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i + w_0 - t_i) = 0$$

The ML estimation for $\beta$ is obtained by

$$-\frac{1}{2} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i + w_0 - t_i)^2 + \frac{n}{2\beta} = 0 \qquad \Longrightarrow \qquad \beta_{ML} = \left( \frac{1}{n} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i + w_0 - t_i)^2 \right)^{-1}$$

## Overfitting

Maximizing the likelihood of the observed dataset tends to result into an estimate too sensitive to the dataset values, hence into overfitting. The obtained estimates are suitable to model observed data, but may be too specialized to be used to model different datasets.

## Penalty functions

An additional function $P(\boldsymbol{\theta})$ can be introduced with the aim to limit overfitting and the overall complexity of the model. This results in the following function to maximize

$$C(\boldsymbol{\theta}|\mathbf{X}) = l(\boldsymbol{\theta}|\mathbf{X}) - P(\boldsymbol{\theta})$$

as a common case, $P(\boldsymbol{\theta}) = \frac{\gamma}{2}\|\boldsymbol{\theta}\|^2$, with $\gamma$ a tuning parameter.

Nel caso $\theta^*_{MAP}$, l'overfitting ha meno effetto: per un po', $\theta^*$ non viene determinato solo sulla base dei dati. Punto di debolezza Bayes: il prior, converrebbe una $p(\theta)$ più neutra possibile ma in fatto tendo conto che serve una distribuzione coniugata. Si cerca di adeguare distrib. coniugate ma con max. varianza possibile.

# Maximum a posteriori estimate


## Idea

Inference through maximum a posteriori (MAP) is similar to ML, but $\boldsymbol{\theta}$ is now considered as a random variable (bayesian approach), whose distribution has to be derived from observations, also taking into account previous knowledge (prior distribution). The parameter value maximizing

$$p(\boldsymbol{\theta}|\mathcal{T}) = \frac{p(\mathcal{T}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{T})}$$

is computed.

## Estimate

$$\boldsymbol{\theta}^*_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\mathcal{T}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{T}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$
$$= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}|\mathcal{T})p(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \left(l(\boldsymbol{\theta}|\mathcal{T}) + \ln p(\boldsymbol{\theta})\right)$$

passiamo anche a $\ln p(\vartheta|Z)$ per

$\Downarrow$

$\ln\left(p(\vartheta|Z)\right)$

which results into

$$\boldsymbol{\theta}^*_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left( \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \right)$$

parte di max. verosimiglianza

or

$$\boldsymbol{\theta}^*_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left( \sum_{i=1}^{n} \ln p(t_i|\mathbf{x}_i, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \right)$$

$\ln$ della distribuzione a priori.

il $+ \ln p(\vartheta)$ ha l'effetto di non far segnare troppo i dati (almeno finché non sono troppi).

## Hypothesis

Assume $\boldsymbol{\theta}$ is distributed around the origin as a multivariate gaussian with uniform variance and null covariance.That is,

$$p(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \sigma^2) = \frac{1}{(2\pi)^{d/2}\sigma^d} e^{-\frac{\|\boldsymbol{\theta}\|^2}{2\sigma^2}} \propto e^{-\frac{\|\boldsymbol{\theta}\|^2}{2\sigma^2}}$$

## Inference

From the hypothesis,

$$\boldsymbol{\theta}^*_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ p(\boldsymbol{\theta}|\mathcal{T}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ (l(\boldsymbol{\theta}|\mathcal{T}) + \ln p(\boldsymbol{\theta}))$$

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left( l(\boldsymbol{\theta}|\mathcal{T}) + \ln e^{-\frac{\|\boldsymbol{\theta}\|^2}{2\sigma^2}} \right) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left( l(\boldsymbol{\theta}|\mathcal{T}) - \frac{\|\boldsymbol{\theta}\|^2}{2\sigma^2} \right)$$

which is equal to the penalty function introduced before, if $\gamma = \frac{1}{\sigma^2}$

# MAP estimate

Approccio applicato al lancio della moneta: qui c'è la distribuzione a priori.

## Example

Collection $\mathbf{X}$ of $n$ binary events, modeled as a Bernoulli distribution with unknown parameter $\phi$. Initial knowledge of $\phi$ is modeled as a Beta distribution:

$$p(\phi|\alpha,\beta) = \text{Beta}(\phi|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\phi^{\alpha-1}(1-\phi)^{\beta-1}$$

Log-likelihood

$$l(\phi|\mathbf{X}) = \sum_{i=1}^{n}(x_i \ln\phi + (1-x_i)\ln(1-\phi)) = n_1 \ln\phi + n_0 \ln(1-\phi)$$

max verosimiglianza

$$\frac{\partial}{\partial\phi}\left(l(\phi|\mathbf{X}) + \ln \text{Beta}(\phi|\alpha,\beta)\right) = \left|\frac{n_1}{\phi} - \frac{n_0}{1-\phi}\right| + \left|\frac{\alpha-1}{\phi} - \frac{\beta-1}{1-\phi}\right| = 0 \implies$$

parte del prior.

$$\phi^*_{MAP} = \frac{N_1+\alpha-1}{n_0+n_1+\alpha+\beta-2} = \frac{n_1+\alpha-1}{n+\alpha+\beta-2}$$

$\alpha, \beta$ del prior, se $n$ diventa grande, parallelamente anche $N_1, N_1$ crescono $\Rightarrow \sim \frac{n_1}{n}$ ed $\alpha, \beta$ non hanno più molto effetto.

## Note

Abbiamo $p(X|\vartheta)$

Come scelgo $\vartheta$:
- Con max. verosimiglianza
- Con MAP

Avrò quindi diversi modelli: $p(X|\vartheta_1)$, $p(X|\vartheta_2)$ ...

ne scelgo 1 (il migliore) e lo uso per le predizioni.

## Gamma function

The function

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

is an extension of the factorial to the real numbers field: in fact, for any integer $x$,

$$\Gamma(x) = (x-1)!$$

Scelgo quindi un $p(X|\vartheta_i)$ e faccio la previsione con lui. Serve un criterio di scelta:

- previsioni precise $\Rightarrow$ max verosimiglianza.
- preferisco con una certa conoscenza pregressa, è MAP. Se un predittore ci prende molto, tenderà ad essere preferito, man mano che aumentano i dati.

- potrei provare a consultare tutti i predittori: se devo fare una certa predizione, vedo quellef atte da tutti i predittori e comprle. Ma i predittori non sono tutti equivalenti (in base alle predizioni passate), quindi occorre pesare le predizioni per poi comporle. La misura viene fissata per la qualità della predizione passata, non è la selezione di un modello ma la comoposizione di più di essi.

- porta lontano da quanto detto fin ora, si può fare in ambito Bayesiano e sarebbe l'approccio "puro", il vero approccio Bayesiano ed è analiticamente un po' più complesso ed in realtà più robusto in un certo senso: scegliendo un solo modello, questo può non avere una buona qualità mentre tenendo conto di quello che dicono tutti i modelli può essere più robusto.

Ho modelli parametrici rispetto a $\theta$ : $p(X|\theta)$ , dato x voglio determinare $p(x)$. Quindi, se ho scelto un solo modello ho $\theta^*_{ML} \Rightarrow x\ p(x|\theta^*_{ML})$ o $p(x|\theta^*_{MAP})$. Qui l'idea è che per arrivare alla stima di prob. siamo passati per la distrib. a posteriori.

$p(\theta|X)$ : è la prob. di $\Theta$, se i dati sono $X$ (e la premessa $p(\theta)$ : se ho $\theta_1$ e $\theta_2$ e $p(\theta_1|x) > p(\theta_2|x)$ $\theta_1$ va meglio di $\theta_2$ per giustificare i dati.

(che il modello sia esattamente quello).

| | |
|---|---|
| $\vartheta_1$ | $p(\vartheta_1 \mid X)$ |
| $\vartheta_2$ | $p(\vartheta_2 \mid X)$ |
| ⋮ | |
| $\vartheta_K$ | $p(\vartheta_K \mid X)$ |

qualità di come $\vartheta_i$ giustifica $X$. Se gli do $x$, ottengo fra una predizione:

$$p(x \mid \vartheta_1), \; p(x \mid \vartheta_2) \dots, \; p(x \mid \vartheta_K).$$

↳ questa è la qualità della predizione.

Allora: $p(x \mid X) = \sum_{i=1}^{K} p(x \mid \vartheta_i) \cdot p(\vartheta_i \mid X)$ pesa quindi la predizione per la qualità.

È una media: $\mathbb{E}_{p(\vartheta_i \mid X)}\, p(x \mid \vartheta_i)$ ma allora non è più parametrico perché ho considerato il valore di tutti i parametri.

Nel continuo: $p(x \mid X) = \int_{\vartheta} p(x \mid \vartheta) \cdot p(\vartheta \mid X) \, d\vartheta$

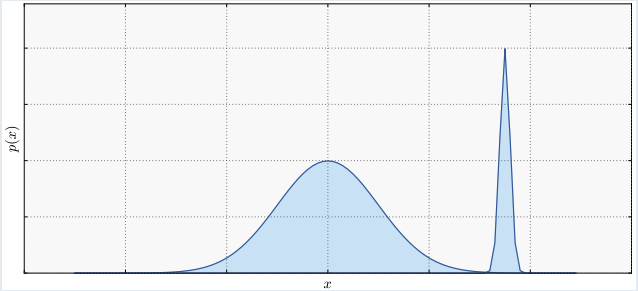
È l'approccio FULLY BAYESIAN ed è l'approccio più elegante, in cui ci si sbilancia di meno ma lo considereremo poco (bisogna fare un'espressione in $x$ è complicata).

## Mode and mean

Once the posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta})d\boldsymbol{\theta}}$$

is available, MAP estimate computes the most probable value (mode) $\boldsymbol{\theta}_{MAP}$ of the distribution. This may lead to inaccurate estimates, as in the figure below:



## Mode and mean

A better estimation can be obtained by applying a fully bayesian approach and referring to the whole posterior distribution, for example by deriving the expectation of $\boldsymbol{\theta}$ w.r.t. $p(\boldsymbol{\theta}|\mathbf{X})$,

$$\boldsymbol{\theta}^* = E_{p(\boldsymbol{\theta}|\mathbf{X})}[\boldsymbol{\theta}] = \int_{\boldsymbol{\theta}} \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta}$$

Model selection: fin ora abbiamo visto come scegliere i migliori parametri, ma a monte c'è la scelta del migliore modello e poi ci sono degli iper-parametri che lo caratterizzano meglio

- es: la classificazione con i polinomi, una volta scelto il grado scelgo i valori ottimi per i coefficienti. TUtto ciò che non ha a che fare con la scelta dei parametri è la model selection

## Example

Collection $\mathbf{X}$ of $n$ binary events, modeled as a Bernoulli distribution with unknown parameter $\phi$. Initial knowledge of $\phi$ is modeled as a Beta distribution:

$$p(\phi|\alpha,\beta) = \text{Beta}(\phi|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\phi^{\alpha-1}(1-\phi)^{\beta-1}$$

Posterior distribution

$$\begin{aligned} p(\phi|\mathbf{X},\alpha,\beta) &= \frac{\prod_{i=1}^{N}\phi^{x_i}(1-\phi)^{1-x_i}\,p(\phi|\alpha,\beta)}{p(\mathbf{X})} \\ &= \frac{\phi^{N_1}(1-\phi)^{N_0}\,\phi^{\alpha-1}(1-\phi)^{\beta-1}}{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\,p(\mathbf{X})} = \frac{\phi^{N_1+\alpha-1}(1-\phi)^{N_0+\beta-1}}{Z} \end{aligned}$$

Hence,

$$p(\phi|\mathbf{X},\alpha,\beta) = \text{Beta}(\phi|\alpha+N_1,\beta+N_0)$$

# Model selection

In the process described, a model (structure, hyper-parameter values) has been identified, in come way. How can we deal with this problem?

This is performed through model selection: identify, in a set of possible models, the one which we expect is best to represent the available data.

Indeed, the one whose best (or a good) instantiation is best to represent the available data

We need a way to compare models (not their instantiations), given the dataset

Model selection: fin ora abbiamo visto come scegliere i migliori parametri, ma a monte c'è la scelta del migliore modello e poi ci sono degli iper-parametri che lo caratterizzano meglio

- es: la classificazione con i polinomi, una volta scelto il grado scelgo i valori ottimi per i coefficienti. TUtto ciò che non ha a che fare con la scelta dei parametri è la model selection

# Theoretical approach

Use the posterior probability of each model, given the dataset

$$p(m|\mathcal{T}) = \frac{p(\mathcal{T}|m)p(m)}{p(\mathcal{T})}$$

Cerco di stimare il miglior modello dato $\mathcal{T}$ come facevo per i parametri.
Ma l'applicazione è più complessa.

Observe that:

- If we assume that no specific knowledge on probabilistic models is initially available, then the prior distribution is uniform.
- The evidence $p(\mathcal{T})$ is a constant with respect to $m$

As a consequence, $p(m|\mathcal{T}) \propto p(\mathcal{T}|m)$ and we may refer to the likelihood $p(\mathcal{T}|m)$ in order to compare models

Prob. tutti il modello (es. un polinomio di grado 7) con certi valori dei coefficienti

The distribution $p(\mathcal{T}|m)$ is also the evidence of the dataset w.r.t. model parameters

$$p(\mathcal{T}|m) = \int_{\boldsymbol{\theta}} p(\mathcal{T}|\boldsymbol{\theta}, m)\, p(\boldsymbol{\theta}|m)\, d\boldsymbol{\theta}$$

C'è un approccio Bayesiano, con cui non si va lontano.

# Model selection in practice

Model selection è empirica: provo a considerare una tipologia di modello, es polinomi. Cerchiamo di apprendere dai dati il miglior polinomio di deg 2. Lo applico su dati e misuro la qualità delle predizioni.

## Validation

**Test set** Dataset is split into Training set (used for learning parameters) and Test set (used for measuring effectiveness). Good for large datasets: otherwise, small resulting training and test set (few data for fitting and validation)

**Cross validation** Dataset partitioned into $K$ equal-sized sets. Iteratively, in $K$ phases, use one set as test set and the union of the other $K-1$ ones as training set ($K$-fold cross validation). Average validation measures.

As a particular case, iteratively leave one element out and use all other points as training set (Leave-one-out cross validation).

Time consuming for large datasets and for models which are costly to fit.

Passo a grado 2, costruisco una griglia di iper-parametri, posso anche cambiare modello.
Confronto i risultati, e' un' indicazione empirica di quel e':

tipo modello – val. iper parametri. (dipende la quantità di fatica la griglia da...)

è una ricerca esaustiva => time consuming.

Fissiamo struttura di modello e valore degli iperparametri, poi guardando ai dati (a partire dal train set) troviamo il valore migliore dei parametri. Prendiamo poi il modello per testarlo sui dati e vedere come si comporta, i dati devono essere diversi quindi sono il validation set (tutti diversi da quelli del train set).

Abbiamo quindi tutti i migliori modelli appresi in questo modo con le relative prestazioni su questo insieme di dati, ne scegliamo uno: se vogliamo prendere il migliore, a noi fa SEMPRE COMODO sapere come si comporterà il modello su dati nuovi.

Serve quindi un 3° insieme di dati, perché i primi due sono entrati in qualche modo nella scelta del modello:
- il training set è l'insieme dei dati che uso, una bolta fissato modello ed iper-parametri, per determinare il valore dei parametri
- validation set è un insieme di dati su cui applico i modelli derivati ottimizzando sul training set

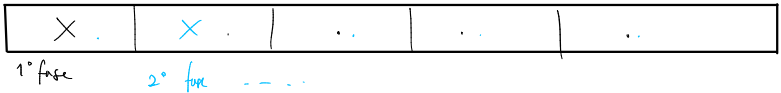
Prendo il modello che si comporta meglio sul validation set, serve quindi un insieme di dati mai utilizzati per capire quale modello sia il migliore. Ci sono quindi 3 set:
- training set, che uso per minimizzare rishcio empirico / massimizzare la verosimiglianza
- validation set che uso per capire il miglior valore degli iper-parametri
- test set: da indicazione di massima di come ci dobbiamo aspettare che il modello si comporti su dati nuovi.

Un approccio è quello di dire che arriva un dataset e lo divido nei 3 insiemi diversi: fisso un valore degli iper-parametri ed uso il validation set per trovare il valore migliore per essi, quindi lo suo come se fosse un test set.

Entra in gioco un'altra considerazione:
- tolto test set e validation set ma rischio che il training set diventi troppo piccolo
- può accadere che a livello train set e test set ho un test set particolare tale per cui il dato magari non è presente nel test set?

Si applica la cross-validation, al costo di un aumento del tempo di apprendimento del modello.

Abbiamo un certo data set (il test set è già tolto):




fisso un valore k, es:5, il processo generale è la k-fold cross validation. Prendo l'intero data set e lo divido in 5 parti uguali, si effettua in 5 step:

1) si prende il primo pezzo e si valuta il modello (validation set), train set sono gli altri 4
2) si prende il secondo pezzo e si usa come validation set e poi quello che resta come train set
i) i-esimo è usato come validation set, i-1 come train set

Le prestazioni vanno mediate: faccio la media delle misure osservate nei 5 casi.

Questo ci permette di confrontare anche modelli che hanno iper-parametri diversi, prendiamo i migliori di quella classe e li confrontiamo sui dati

## Information measures

Faster methods to compare model effectiveness, based on computing measures which take into account data fitting and model complexity.

Akaike Information Criterion (AIC) Let $\boldsymbol{\theta}$ be the set of parameters of the model and let $\boldsymbol{\theta}_{ML}$ be their maximum likelihood estimate on the dataset $\mathbf{X}$. Then,

$$AIC = 2|\boldsymbol{\theta}| - 2\log p(\mathbf{X}|\boldsymbol{\theta}_{ML}) = 2|\boldsymbol{\theta}| - 2\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{X})$$

lower values correspond to models to be preferred.

Bayesian Information Criterion (BIC) A variant of the above, defined as

$$\begin{aligned} BIC &= |\boldsymbol{\theta}| - \log|\mathbf{X}|2\log p(\mathbf{X}|\boldsymbol{\theta}_{ML}) \\ &= |\boldsymbol{\theta}|\log|\mathbf{X}| - 2\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{X}) \end{aligned}$$

Indici statistici che danno la qualità del modello, data la sua complessità.
Introducono dei correttivi per considerare la cardinalità del n° di parametri.

Vogliamo vedere come si csotruisce un classificatore, ci concentriamo su quelli Bayesiani.

Classificatore Bayesiano: abbiamo due classi, $C_0$ e $C_1$, cerchiamo di stiamre la probabilità della classe $C_1$ dato x $p(C_1|x)$. Data una stima di questa probabilità possiamo assegnare x a $C_0$ o $C_1$ data questo valore di probabilità (ci può sempre essere la soglia a valle).

Quindi: $C_0$, $C_1$ $x \in C_0$? $x \in C_1$? L'idea è ottenere $p(C_1|x)$ e $p(C_0|x)$ } li confrontiamo:

se $p(C_1|x) > p(C_0|x) \Rightarrow x \in C_1$, altrimenti $x \in C_0$

C'è il thm. di Bayes: $p(C_1|x) = \frac{p(x|C_1) \cdot p(C_1)}{p(x)}$ e $p(C_0|x) = \frac{p(x|C_0) \cdot p(C_0)}{p(x)}$

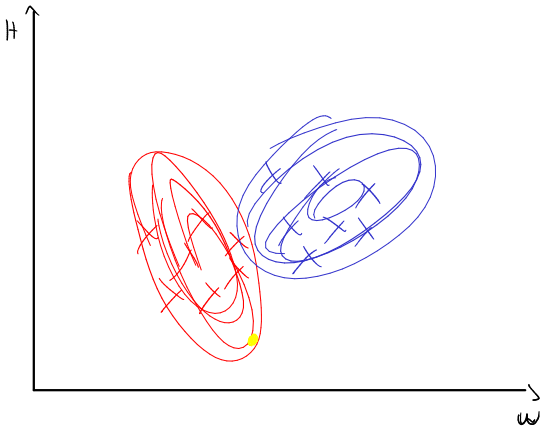
$\Rightarrow$ vado a vedere se $p(x|C_1) \cdot p(C_1) > p(x|C_0) \cdot p(C_0)$

$p(C_1|x)$: prob. condizionata a posteriori che l'elemento $x \in C_1$ nell'ipotesi che l'elemento sia proprio x. Prendo x a caso, che prob ho che $\in C_1$? $p(C_0|x)$ è il complementare. In Bayes abbiamo a priori ($p(x|C_1)$).

Come interpreto $p(x|C_1)$

Assumo di stare in $C_1$, qual è la prob. che un elemento estratto a caso sia $x$? È come se fosse la distribuzione di $C_1$. Stesso per $C_0$: ho due modelli probabilistici di $C_0$ e $C_1$

supponiamo che tutti gli elementi siano valori in uno spazio a 2D (sesso altezza) e di avere un target (MF). Allora $p(x|M)$ è la distribuzione di probabilità di peso ed altezza fra i maschi e $p(x|F)$ è la stessa ma tra le femmine. Quindi, supponiamo che il nostro training set vada così:




M

F

Avrò:

$\mu_M$ $\mu_F$

$\sigma_M$ $\sigma_F$

suppongo che la distribuzione delle femmine sia una Gaussiana cerco la migliore Gaussiana ed ottengo un mio modello di $p(x|F)$ . Per i maschi potrei fare la stessa cosa, ottenendo $p(x|M)$.

Considerando tutti gli elementi insieme potrei avere una Gaussiana che ci da p(x), ma non ci interessa

Avendo le due espressioni analitiche $p(x|M)$ e $p(x|F)$ posso prendere un punto avrà due probabilità (una in quanto M ed una in quanto F)

Se quindi ho un valore di x posso determinare $p(x|C_0)$ e $p(x|C_1)$ : posso raffrontare le due probabilità tenendo conto anche di quelle a priori:
- se la probabilità delle femmine è più alta ma queste sono di meno nel training set, allora ho una probabilità a priori, il raffronto fra le due probabilità che sono le verosimiglianze tiene conto del prior.

Abbiamo qundi sia una conoscenza pregressa che la verosimiglianza e questo viene tenuto in conto dalla formula di Bayes.

Stimiamo $p(x|C_1)$ facendo fitting: consideriamo la migliore Gaussiana che rappresenta M ed F e poi le raffrontiamo. Cerchiamo la migliore distribuzione di una classe, esempio Gaussiana, che rappresenti quei dati. In questo passo stiamo facendo apprendimento non supervised: ho solo maschi (o femmine) e voglio rappresentarlo nel modo migliore possibile con un modello probabilistico. Prendiamo la migliore Gaussiana (cambia media e matrice di cov).

Servono anche $p(C_1)$ e $p(C_0)$: per stimarle, uso nuovamente la massima verosimiglianza (come il caso del lancio della moneta, vedo i valori nel dataset, ricorda il modello di Bernoulli).

Abbiamo una scelta arbitraria: scegliamo la distribuzione, esempio Gaussiana, va bene se i dati si rappresentano bene con questa distribuzione. Se invece la massima vero simiglianza è limitata, ritorna il discorso della model selection, per cui potrei cambiare la distribuzione. Potrei anche ad esempio applicare un test di Gaussianità. La filosofia è sempre la stessa:
- classe di modelli, trovo quello migliore. Nel caso non supervised il migliore che descrive i dati, caso supervisionato il milgiore per predire i dati

Quindi per determinare se una persona 1.62mx52kg è M o F, vedo le probabilità dei due universi, le raffronto tenendo conto che un elemento preso a caso sia femmina o maschio. Confronto quindi:

$p(M|x)$ e $p(F|x)$ ATTENZIONE: $p(x|M) \neq p(x|M)$

Partono quindi da una descrizione di come sono fatte le classi.

# Language modeling

A language model is a (categorical) probability distribution on a vocabulary of terms (possibly, all words which occur in a large collection of documents).

## Use

A language model can be applied to predict the next term occurring in a text. The probability of occurrence of a term is related to its information content and is at the basis of a number of information retrieval techniques.

## Hypothesis

It is assumed that the probability of occurrence of a term is independent from the preceding terms in a text (bag of words model).

## Generative model

Given a language model, it is possible to sample from the distribution to generate random documents statistically equivalent to the documents in the collection used to derive the model.

Vogliamo classificare un documento, indicando con 0 se un termine compare in un documento ed 1 se compare (è il vettore caratteristico). Non tengo conto di diversi fattori:
  - ordine dei termini
  - molteplicità dei termini.
Le feature in questo caso sono i termini del dizionario, quindi ho tante feature quanti sono i termini del dizionario, il valore che una feature assume in un documento è 0,1. Ho tante feature con pochi valori, anche se cambio i valori delle feature non cambia molto, un documento == ad un vettore, lungo perché sono tutti i termini del dizionario, quindi a differenza del caso M/F siamo in uno spazio a molte dimensioni.

L'approccio usato su M/F ci piace, vogliamo fare un classificatore binario Bayesiano su documenti: sentiment analysis, tweet su twitter. Il documento esprime un sentimento negativo, o positivo. Nel nostro train set abbiamo un insieme di documenti e il target è offensivo/non offensivo. Altro caso è il filtro anti-spam, per determinare si guarda il contenuto del messaggio ed i termini che vi compaiono. Il classificatore quindi, dato un messaggio, deve dire qual è la probabilità che il messaggio sia spam e qual è quella che sia un messaggio significativo. È un classificatore binario.

In un mail client viene detto di controllare lo spam, perché può capitare che un messaggio sia classificato male, o viceversa per indicare che un messaggio "buono" sia dato come spam. Facendo ciò, il training set che il classificatore ha a disposizione aumenta, perché stiamo dando un nuovo elemento con target definito. Ogni indicazione permette quindi al classificatore di migliorare se stesso.

Supponiamo che l'antispam sia un classificatore Bayesiano, vogliamo seguire lo stesso approccio di prima

Dato $x$, documento e' del $(0, \dots 0, \dots 1, \dots 0)$ vogliamo stimare una p!

$(x_1, x_2 \dots \dots x_{|V|})$

$p(S \mid [0, 0, \dots 1, \dots 0])$ spam e lo raffronto a $p(M \mid [0, 0, \dots \dots 1, \dots 0])$

$p(M \mid [\dots]) \propto p([0, 0010 \dots] \mid M)\, p(M)$ e lo raffronto con

$p(S \mid [\dots]) \propto p([0, 0010 \dots] \mid M)\, p(S)$

Per $p(M)$ e $p(S)$ vale quanto detto prima. Abbiamo poi le due probabilità condizionate: se avessimo un solo termine, potremmo usare la distrib. di Bernoulli, ma ho un vettore di 0,1. Faccio riferimento alla distribuzione categorica

N e' il numero di termini

# Language model

il passo successivo consiste nel considerare, qual è la prob che sia il 1° termine, ed il 2°, ed il 3° ...?

Se $N = 5$ (palla, pippo, luna, terra, mare) e chiamo le prob. $\phi_1, \phi_2, \phi_3, \phi_4, \phi_5$ ($\sum \phi_i = 1$)

- Let $\mathcal{T} = \{t_1, \dots, t_n\}$ be the set of terms occurring in a given collection $\mathcal{C}$ of documents, after stop word (common, non informative terms) removal and stemming (reduction of words to their basic form).
- For each $i = 1, \dots, n$ let $m_i$ be the multiplicity (number of occurrences) of term $t_i$ in $\mathcal{C}$
- A language model can be derived as a categorical distribution associated to a vector $\hat{\boldsymbol{\phi}} = (\hat{\phi}_1, \dots, \hat{\phi}_n)^T$ of probabilities: that is,

$$0 \leq \hat{\phi}_i \leq 1 \quad i = 1, \dots, n \qquad\qquad \sum_{i=1}^{n} \hat{\phi}_i = 1$$

where $\hat{\phi}_j = p(t_j|\mathcal{C})$

Posso definire $(\phi_1, \phi_2, \dots \phi_n)$ che rappresentano la prob. che il termine estratto sia il 1°, il 2° etc... $\phi_j = p(t_j)$

Questo è il MODELLO DI LINGUAGGIO

Per scegliere il migliore fra i modelli di linguaggio è fare quanto segue:
- ho tutti i documenti, spam / no spam
- $\phi_1$ guardo tutti gli spam e voglio determinare per tutti i documenti spam una probabilità del primo termine, per il termine "palla".

Stimo quindi phi_1 che è : $\frac{\#t_1}{\#t_1+\#t_2+\#t_3+\#t_4+\#t_5}$ che è la stima di massima verosimiglianza fatta fino ad ora. (in Bernoulli erano 2). Determino quindi Spam $[\hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3, \hat{\phi}_4, \hat{\phi}_5]$ che può essere:
- 35% palla
- 10% pippo
- 5% luna
- 25% terra
- 25% mare

ottenendo il vettore di lunghezza 5. Ora, preso un termine a caso secondo quello che vedo su come sono fatti i documenti spam ho una certa probabilità che il termina sia palla, pippo etc ...

Consideriamo quindi un problema leggermente più semplice: siamo partiti dal dire che abbiamo un messaggio e voglimao determinare se sia spam / non spam. Sotto-problema: il messaggio ha solo un termine, qual è la probabilità che il messaggio sia spam e quale quella che sia non spam. Non vedo il documento, so che compare il termine "terra", quali sono le due probabilità.

$p(S \mid \text{"terra"})$ Confronto con $p(M \mid \text{"terra"})$

$p(\text{"terra"} \mid S)\ p(S)$ $\quad$ $p(\text{"terra"} \mid M)\ p(M)$

$\hat{\phi}_4^S$ $\quad$ $\hat{\phi}_4^M$

Va fatto poi su un vettore, va adattato al caso in cui ho insiemi di termini.

# Learning a language model by ML

Come apprendiamo il modello di linguaggio:

Applying maximum likelihood to derive term probabilities in the language model results into setting

$$\hat{\phi}_j = p(t_j|\mathscr{C}) = \frac{m_j}{\sum_{k=1}^{n} m_k} = \frac{m_j}{N}$$

vettore di documento

where $N = \sum_{i=1}^{n} m_i$ is the overall number of occurrences in $\mathscr{C}$ after stopword removal.

## Smoothing

According to this estimate, a term $t$ which never occurred in $\mathscr{C}$ has zero probability to be observed (black swan paradox). Due to overfitting the model to the observed data, typical of ML estimation.

Solution: assign small, non zero, probability to events (terms) not observed up to now. This is called smoothing.

Se un termine del dizionario è "nero" e non compare mai ⇒ la prob. attribuita è 0 ed è pericoloso in statistica (il paradosso del cigno nero). Evento impossibile ≠ mai osservato, quindi prob. 0 non si attribuisce mai all'evento, ma se ne assegna una bassa ⇒ è lo smoothing.

Serve un metodo per entrare prob. di sugli eventi

We may apply the dirichlet-multinomial model:

- this implies defining a Dirichlet prior $\mathrm{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha})$, with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)$ that is,

$$p(\phi_1, \dots, \phi_n|\boldsymbol{\alpha}) = \frac{1}{\Delta(\alpha_1, \dots, \alpha_n)} \prod_{i=1}^{n} \phi_i^{\alpha_i - 1}$$

- the posterior distribution of $\boldsymbol{\phi}$ after $\mathscr{C}$ has been observed is then $\mathrm{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha}')$, where

$$\boldsymbol{\alpha}' = (\alpha_1 + m_1, \alpha_2 + m_2, \dots, \alpha_n + m_n)$$

that is,

$$p(\phi_1, \dots, \phi_n|\boldsymbol{\alpha}') = \frac{1}{\Delta(\alpha_1 + m_1, \dots, \alpha_n + m_n)} \prod_{i=1}^{n} \phi_i^{\alpha_i + m_i - 1}$$

Con l'apporoccio Bayesiano, possiamo assumere una probabilità a priori dove tutti i termini hanno la stessa probabilità di occorrere. All'aumentare del numero di dati che dicono che un certo elemento non occorre, la probabilità di occorrenza diventerà sempre più piccolo ma mai 0, perché c'è l'effetto della distribuzione a priori rispetto a quella a posteriori

Riusciamo a non guardare solo ai dati, introduciamo anche alto. Con l'approccio Bayesiano, il modello di riferimento che ho: il ruolo della distribuzione di Bernoulli viene preso dalla distribuzione multi-nomiale, dove il singolo evento è un insieme di eventi: es. lancio un dado n volte, qual è la probabilità che n_1 volte esca 1, … , n_6 volte esca 6, fissati gli n_i), è un estensione della Bernoulliana

# Bayesian learning of a language model

The language model $\hat{\boldsymbol{\phi}}$ corresponds to the predictive posterior distribution

$$\hat{\phi}_j = p(t_j|\mathscr{C}, \boldsymbol{\alpha}) = \int p(t_j|\boldsymbol{\phi}) p(\boldsymbol{\phi}|\mathscr{C}, \boldsymbol{\alpha}) d\boldsymbol{\phi}$$
$$= \int \phi_j \text{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha}') d\boldsymbol{\phi} = E[\phi_j]$$

where $E[\phi_j]$ is taken w.r.t. the distribution $\text{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha}')$. Then,

$$\hat{\phi}_j = \frac{\alpha'_j}{\sum_{k=1}^n \alpha'_k} = \frac{\alpha_j + m_j}{\sum_{k=1}^n (\alpha_k + m_k)} = \frac{\alpha_j + m_j}{\alpha_0 + N}$$

quante volte for assumato il termine nei documenti

numero documenti.

The $\alpha_j$ term makes it impossible to obtain zero probabilities (Dirichlet smoothing).

Non informative prior: $\alpha_i = \alpha$ for all $i$, which results into

$$p(t_j|\mathscr{C}, \boldsymbol{\alpha}) = \frac{m_j + \alpha}{\alpha V + N}$$

where $V$ is the vocabulary size.

$\alpha_j$ non è altro che l'ipotesi del prior: assumiamo di aver visto già per $\alpha_j$ volte l'occorrenza del valore j.

Assumiamo nel prior di aver visto nel passato lo stesso numero di occorrenze per tutti i termini, questo fa si che non avrò mai il valore 0 per $\phi_j$

Anche se N diventa tanto grande, perché ho tante osservazioni e quindi anche il numero di occorrenze del termine cresce, il rapporto tende asintoticamente a $\frac{m_j}{N}$ . Così come se $m_j$ tende a 0 quello tenderà a 0 ma non sarà mai pari a 0

Dobbiamo quindi selezionare la miglior distribuzione categorica ragionando in modo naive, ma stando attenti a cose che possono non piacerci, come probabilità 0. Abbiamo comunque definito un nostro modello di linguaggio:

$S_{pam} \left( \hat{\phi}_1^s, \hat{\phi}_2^s, \ldots, \hat{\phi}_M^s \right)$ $M = |V|$

$N_{on} S_{pam} \left( \hat{\phi}_1^n, \hat{\phi}_2^n, \ldots, \hat{\phi}_M^n \right)$

A language model can be applied to derive document classifiers into two or more classes.

- given two classes $C_1, C_2$, assume that, for any document $d$, the probabilities $p(C_1|d)$ and $p(C_2|d)$ are known: then, $d$ can be assigned to the class with higher probability
- how to derive $p(C_k|d)$ for any document, given a collection $\mathscr{C}_1$ of documents known to belong to $C_1$ and a similar collection $\mathscr{C}_2$ for $C_2$? Apply Bayes' rule:

$$p(C_k|d) \propto p(d|C_k)p(C_k)$$

  the evidence $p(d)$ is the same for both classes, and can be ignored.
- we have still the problem of computing $p(C_k)$ and $p(d|C_k)$ from $\mathscr{C}_1$ and $\mathscr{C}_2$

## Computing $p(C_k)$

The prior probabilities $p(C_k)$ $(k = 1, 2)$ can be easily estimated from $\mathscr{C}_1, \mathscr{C}_2$: for example, by applying ML, we obtain

$$p(C_k) = \frac{|\mathscr{C}_1|}{|\mathscr{C}_1| + |\mathscr{C}_2|}$$

può capitare che nel dataset una delle due classi non capita mai. Ritorna quindi lo smoothing, ma assumiamo il dataset grande $\Rightarrow$ entrambe le classi

## Computing $p(d|C_k)$

For what concerns the likelihoods $p(d|C_k)$ $(k = 1, 2)$, we observe that $d$ can be seen, according to the bag of words assumption, as a multiset of $n_d$ terms

$$d = \{\bar{t}_1, \bar{t}_2, \ldots, \bar{t}_{n_d}\}$$

By applying the product rule, it results

$$\begin{aligned} p(d|C_k) &= p(\bar{t}_1, \ldots, \bar{t}_{n_d}|C_k) \\ &= p(\bar{t}_1|C_k)p(\bar{t}_2|\bar{t}_1, C_k) \cdots p(\bar{t}_{n_d}|\bar{t}_1, \ldots, \bar{t}_{n_d-1}, C_k) \end{aligned}$$

Come stimo la probabilità congiunta: posso vedere lo spazio dei possibili eventi come tutti i possibili vettori, ma avrebbe una cardinalità spropositata, perché considero tutti i vettori di lunghezza n

Geometricamente: ho un insieme di vettori di dimensione elevata, es: 10000 e magari sono 5000 e quindi sono elementi molto lontani fra loro. Non ci faccio molto da un punto di vista probabilistico, dovrei avere molti punti. Lo spazio in cui mi muovo è troppo grande, in ML questo prende il nome di MALEDIZIONE DELLA DIMENSIONALITÀ (qualcosa del genere): sta a rappresentare una sitauzione in cui i dati sono vettori di dimensione molto grande e mi ci muovo con un numero di dati limitato, che dovrebbe crescere esponenzialmente.

Essendo i vettori lunghi, nell'insieme dei possibili eventi li ho molto sparsi per cui occorre semplificare in qualche modo:

$$p(x_1, x_2, x_3) = p(x_1, x_2 | x_3) \cdot p(x_3) = p(x_1 | x_2, x_3) \cdot p(x_2 | x_3) p(x_3) \dots$$

è un'identità. Ora faccio un'ipotesi semplificativa:

$$p(t_2 | t_1, C_k) = p(t_2 | C_k)$$, quindi es $p(x_2 | x_3) = p(x_2)$. quindi che gli eventi sono indipendenti. $\Rightarrow p(t_1 | C_k)\, p(t_2 | C_k) \cdots p(t_{n_d} | C_k)$. Quindi, l'occorrenza di un termine non dà informazioni sull'occorrenza di un altro. L'indipendenza è però condizionata: una volta che so se il documento è S / NS, considerato che un certo termine è compreso questo non influenza l'occorrenza di altri termini.

L'ipotesi naive è quindi l'indipendenza per due qualunque termini.

## The naive Bayes assumption

Computing $p(d|C_k)$ is much easier if we assume that terms are pairwise conditionally independent, given the class $C_k$, that is, for $i, j = 1. \ldots, n_d$ and $k = 1, 2$,

$$p(\bar{t}_i, \bar{t}_j|C_k) = p(\bar{t}_i|C_k)p(\bar{t}_2|C_k)$$

as, a consequence,

$$p(d|C_k) = \prod_{j=1}^{n_d} p(\bar{t}_j|C_k)$$

## Language models and NB classifiers

The probabilities $p(\bar{t}_j|C_k)$ are available for all terms if language models have been derived for $C_1$ and $C_2$, respectively from documents in $\mathscr{C}_1$ and $\mathscr{C}_2$.

# Feature selection by mutual information

## Feature selection

The set of probabilities in a language model can be exploited to identify the most relevant terms for classification, that is terms whose presence or absence in a document best characterizes the class of the document.

## Mutual information

To measure relevance, we can apply the set of mutual informations $\{I_1, \dots, I_n\}$

Deriva dalla teoria dell'informazione.

mutua informazione di $J$ rispetto alla Classe

$$I_j = \sum_{k=1,2} p(t_j, C_k) \log \frac{p(t_j, C_k)}{p(t_j)p(C_k)}$$

$$= \sum_{k=1,2} p(C_k|t_j)p(t_j) \log \frac{p(C_k|t_j)}{p(C_k)} = p(t_j) KL(p(C_k|t_j) \| p(C_k))$$

here, $KL$ is a measure of the amount of information on class distributions provided by the presence of $t_j$. This amount is weighted by the probability of occurrence of $t_j$.

Supponiamo di avere incertezza sul fatto che il documento sia della prima o della seconda classe, questo è legato al prior, che sarà 0.5 / 0.5

Supponiamo che veniamo a conoscere del documento che ci compaia un certo termine $t_j$: sappiamo ora che compare (o non compare) tale termine, quindi abbiamo una informazione in più. Quello che ci chiediamo è: il fatto di avere questa informazione in più è utile a diminuire l'incertezza sulla classe a cui appartiene tale documento?

- se rimane la stessa, aver saputo che il termine compare o non compare non è servito, la feature non mi dice nulla (rimane 0.5 / 0.5)
- se invece l'incertezza diminuisce molto, allora il termine è informativo rispetto alla classe, allora è una feature rilevante

Questa è l'idea della mutua informazione: confronta la probabilità congiunta per le varie classi, dell'elemento con la classe, col prodotto delle porbabilità:

- se le due sono molto simili, allora le cose sono indipendenti. Ho poca informazione, che avviene quando la probabilità congiunta è il prodotto delle probabilità

  Se $t_j$ e $C_k$ sono fra loro indipendenti, vuol dire che osservandolo non da informazione. L'informazione mutua misura questo

È una misura applicabile date due variabili aleatorie, nel nostro caso dato un termine e la classe. Più il valore $I_j$ è elevata, maggiore sarà la diminuzione dell'incertezza sulla classe. Funzionerebbe anche fra due

termini: il fatto di sapere che compare un termine diminuisce l'incertezza di un altro.
La mutua informazione può essere calcolata per ogni termine, per poi ordinare i termini in base a questo valore: più è alto, maggiore è l'indicazione che siamo lontani dall'indipendenza e più è bassa e più siamo vicini all'indipendenza.

## Mutual information

Since $p(t_j, C_k) = p(C_k|t_j)p(t_j) = p(t_j|C_k)p(C_k)$, $I_j$ can be estimated as

$$I_j = p(t_j|C_1)p(C_1)\log\frac{p(t_j|C_1)}{p(t_j)} + p(t_j|C_2)p(C_2)\log\frac{p(t_j|C_2)}{p(t_j)}$$

$$= \phi_{j1}\pi_1 \log\frac{\phi_{j1}}{\phi_{j1}\pi_1 + \phi_{j2}\pi_2} + \phi_{j2}\pi_2 \log\frac{\phi_{j2}}{\phi_{j1}\pi_1 + \phi_{j2}\pi_2}$$

where $\phi_{jk}$ is the estimated probability of $t_j$ in documents of class $C_k$ and $\pi_k$ is the estimated probability of a document of class $C_k$ in the collection.

A selection of the most significant terms can be performed by selecting the set of terms with highest mutual information $I_j$.

Scegliamo quindi i valori più significativi selezionando le colonne (da 600 a 15 per il notebook meno buggoso)
↳ è come model selection, uso il numero che reputo come migliore.