Model inference

Course of Machine Learning
Master Degree in Computer Science
University of Rome "Tor Vergata"
a.a. 2019-2020

Giorgio Gambosi

Purpose
Inferring a *probabilistic model* from a collection of observed data $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. A probabilistic model is a probability distribution over the data domain.

Dataset
A dataset $\mathbf{X}$ is a collection of N observed data, independent and identically distributed (iid): they can be seen as realizations of a single random variable.

Problems considered
Inference objectives:

Model selection Selecting the probabilistic model $\mathcal{M}$ best suited for a given data collection

Estimation Estimate the values of the set $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)$ of parameters of a given model type (probability distribution), which best model the observed data $\mathbf{X}$

Prediction Compute the probability $p(x|\mathbf{X})$ of a new observation from the set of already observed data

Context
Model space $\mathcal{M}$: a model $m \in \mathcal{M}$ is a probability distribution $p(\mathbf{x}|m)$ over data.
Let $p(m)$ be any *prior distribution* of models

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

The corresponding predictive distribution of data is

$$p(\mathbf{x}) = \sum_{m \in \mathcal{M}} p(\mathbf{x}|m)p(m)$$

After the observation of a dataset $\mathbf{X}$, the updated probabilities are

$$p(m|\mathbf{X}) = \frac{p(m)p(\mathbf{X}|m)}{p(\mathbf{X})} \propto p(m)p(\mathbf{X}|m) = p(m)\prod_{i=1}^{n} p(x_i|m)$$

and the predictive distribution is

$$p(\mathbf{x}|\mathbf{X}) = \sum_{m \in \mathcal{M}} p(\mathbf{x}|m)p(m|\mathbf{X})$$

Parametric models
Models are defined as parametric probability distributions, with parameters $\boldsymbol{\theta}$ ranging on a *parameter space* $\boldsymbol{\Theta}$.

A prior parameter distribution $p(\boldsymbol{\theta}|m)$ is defined for a model. The prior predictive distribution is then

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

Posterior parameter distribution
Given a model $m \in \mathcal{M}$, Bayes' formula makes it possible to infer the posterior distribution of parameters, given the dataset $\mathbf{X}$

$$p(\boldsymbol{\theta}|\mathbf{X}, m) = \frac{p(\boldsymbol{\theta}|m)p(\mathbf{X}|\boldsymbol{\theta}, m)}{p(\mathbf{X}|m)} \propto p(\boldsymbol{\theta}|m)p(\mathbf{X}|\boldsymbol{\theta}, m)$$

The posterior predictive distribution, given the model, is

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

According to the bayesian approach to inference, parameters are considered as random variables, whose distributions have to be inferred from observed data.

The approach relies on Bayes' classic result:

Theorem 1 (Bayes)**.** *Let $\mathbf{X}, \mathbf{Y}$ be a pair of (sets of) random variables. Then,*

$$p(\mathbf{Y}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y})}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y})}{\int_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z})d\mathbf{Z}}$$

Motivation
Given a model m, the bayesian approach is aimed to derive the posterior distribution of the set of parameters $\boldsymbol{\theta}$. This requires computing

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

and

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

This is usually impossible to be done efficiently.

Idea
Only an estimate of the "best" value $\hat{\boldsymbol{\theta}}$ in $\boldsymbol{\theta}$ (according to some measure) is performed. The posterior predictive distribution can then be approximated as follows

$$\begin{aligned} p(\mathbf{x}|\mathbf{X}) &= \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{X})d\boldsymbol{\theta} \approx \int_{\boldsymbol{\theta}} p(\mathbf{x}|\hat{\boldsymbol{\theta}})p(\boldsymbol{\theta}|\mathbf{X})d\boldsymbol{\theta} \\ &= p(\mathbf{x}|\hat{\boldsymbol{\theta}}) \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X})d\boldsymbol{\theta} = p(\mathbf{x}|\hat{\boldsymbol{\theta}}) \end{aligned}$$

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

$$L(\boldsymbol{\theta}|\mathbf{X}) = p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta})$$

Log-likelihood

$$l(\boldsymbol{\theta}|\mathbf{X}) = \ln L(\boldsymbol{\theta}|\mathbf{X}) = \sum_{i=1}^{N} \ln p(\mathbf{x}_i|\boldsymbol{\theta})$$

is usually preferrable. The maximum occurs at the same point: $\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, l(\boldsymbol{\theta}|\mathbf{X}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}|\mathbf{X})$

Estimate

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

Solution
Solve the system

$$\frac{\partial l(\boldsymbol{\theta}|\mathbf{X})}{\partial \theta_i} = 0 \qquad\qquad i = 1, \ldots, D$$

more concisely,

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

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

$$\begin{aligned} p(\mathbf{x}|\mathbf{X}) &= \int_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta} \approx \int_{\boldsymbol{\theta}} p(\mathbf{x}|\hat{\boldsymbol{\theta}}_{ML}) p(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta} \\ &= p(\mathbf{x}|\hat{\boldsymbol{\theta}}_{ML}) \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta} = p(\mathbf{x}|\hat{\boldsymbol{\theta}}_{ML}) \end{aligned}$$

Example 2. Collection $\mathbf{X}$ of n binary events, modeled through a Bernoulli distribution with unknown parameter ϕ

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

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)

$$\frac{\partial l(\phi|\mathbf{X})}{\partial \phi} = \frac{N_1}{\phi} - \frac{N_0}{1-\phi} = 0 \qquad \Longrightarrow \qquad \hat{\phi}_{ML} = \frac{N_1}{N_0 + N_1} = \frac{N_1}{N}$$

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 γ a *tuning* parameter.

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

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

is computed.

Estimate

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{MAP} &= \operatorname*{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}) = \operatorname*{argmax}_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{X})p(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \left(l(\boldsymbol{\theta}|\mathbf{X}) + \ln p(\boldsymbol{\theta})\right) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \left(\sum_{i=1}^{N} \ln p(\mathbf{x}_i|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})\right)\end{aligned}$$

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

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

Inference
From the hypothesis,

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{MAP} &= \operatorname*{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \left(l(\boldsymbol{\theta}|\mathbf{X}) + \ln p(\boldsymbol{\theta})\right) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \left(l(\boldsymbol{\theta}|\mathbf{X}) + \ln \exp\left(-\frac{\|\boldsymbol{\theta}\|^2}{2\sigma^2}\right)\right) = \operatorname*{argmax}_{\boldsymbol{\theta}} \left(l(\boldsymbol{\theta}|\mathbf{X}) - \frac{\|\boldsymbol{\theta}\|^2}{2\sigma^2}\right)\end{aligned}$$

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

Example 3. Collection $\mathbf{X}$ of n binary events, modeled as a Bernoulli distribution with unknown parameter ϕ. Initial knowledge of ϕ 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

$$l(\phi|\mathbf{X}) = \sum_{i=1}^{N} (x_i \ln \phi + (1 - x_i) \ln(1 - \phi)) = N_1 \ln \phi + N_0 \ln(1 - \phi)$$

$$\frac{\partial}{\partial \phi} l(\phi|\mathbf{X}) + \ln \text{Beta}(\phi|\alpha, \beta) = \frac{N_1}{\phi} - \frac{N_0}{1 - \phi} + \frac{\alpha - 1}{\phi} - \frac{\beta - 1}{1 - \phi} = 0 \qquad \Longrightarrow$$

$$\hat{\phi}_{MAP} = \frac{N_1 + \alpha - 1}{N_0 + N_1 + \alpha + \beta - 2} = \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2}$$

Gamma function
The function

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

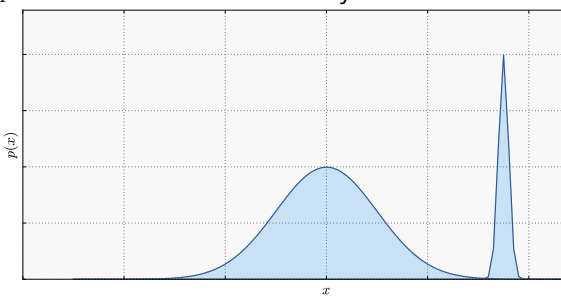
is an extension of the factorial to the real numbers field: hence, for any integer x,

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

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. $p(\boldsymbol{\theta}|\mathbf{X})$,

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

Example 4. Collection $\mathbf{X}$ of n binary events, modeled as a Bernoulli distribution with unknown parameter ϕ. Initial knowledge of ϕ 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

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

Hence,

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

Comparing different models
Let $\mathcal{M}_1, \ldots, \mathcal{M}_m$ be a set of model types, each with its own set of parameters. Given a dataset $\mathbf{X}$, we wish to select the model type which best represents $\mathbf{X}$.

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

$$p(\mathcal{M}_i|\mathbf{X}) = \frac{p(\mathbf{X}|\mathcal{M}_i)p(\mathcal{M}_i)}{p(\mathbf{X})} \propto p(\mathbf{X}|\mathcal{M}_i)p(\mathcal{M}_i)$$

If we assume that no specific knowledge on model types is initially available, then the prior distribution is uniform: as a consequence, $p(\mathcal{M}_i|\mathbf{X}) \propto p(\mathbf{X}|\mathcal{M}_i)$.

Evidence
The distribution $p(\mathbf{X}|\mathcal{M}_i)$ is the evidence of the dataset w.r.t. a model type. It can be obtained by marginalization of model parameters

$$p(\mathbf{X}|\mathcal{M}_i) = \int_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta}, \mathcal{M}_i)p(\boldsymbol{\theta}|\mathcal{M}_i)d\boldsymbol{\theta}$$

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

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

1 Example: learning in the dirichlet-multinomial model

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.

- Let $\mathcal{T} = \{t_1, \ldots, t_n\}$ 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 $i = 1, \ldots, n$ 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 $\hat{\boldsymbol{\phi}} = (\hat{\phi}_1, \ldots, \hat{\phi}_n)^T$ of probabilities: that is,

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

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

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

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

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

Smoothing
According to this estimate, a term t which never occurred in $\mathcal{C}$ 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*.

We may apply the dirichlet-multinomial model:

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

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

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

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

  that is,

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

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

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

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

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

The α_j term makes it impossible to obtain zero probabilities (*Dirichlet smoothing*).

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

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

where V is the vocabulary size.

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 $\mathcal{C}_1$ of documents known to belong to C_1 and a similar collection $\mathcal{C}_2$ 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 $\mathcal{C}_1$ and $\mathcal{C}_2$

Computing $p(C_k)$

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

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

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

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

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 $\mathcal{C}_1$ and $\mathcal{C}_2$.

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 $\{I_1, \ldots, I_n\}$

$$\begin{aligned} I_j &= \sum_{k=1,2} p(t_j, C_k) \log \frac{p(t_j, C_k)}{p(t_j)p(C_k)} \\ &= \sum_{k=1,2} p(C_k|t_j)p(t_j) \log \frac{p(C_k|t_j)}{p(C_k)} = p(t_j)KL(p(C_k|t_j)||p(C_k)) \end{aligned}$$

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.

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

$$\begin{aligned} I_j &= p(t_j|C_1)p(C_1) \log \frac{p(t_j|C_1)}{p(t_j)} + p(t_j|C_2)p(C_2) \log \frac{p(t_j|C_2)}{p(t_j)} \\ &= \phi_{j1}\pi_1 \log \frac{\phi_{j1}}{\phi_{j1}\pi_1 + \phi_{j2}\pi_2} + \phi_{j2}\pi_2 \log \frac{\phi_{j2}}{\phi_{j1}\pi_1 + \phi_{j2}\pi_2} \end{aligned}$$

where ϕ_{jk} is the estimated probability of t_j in documents of class C_k and π_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.