Model inference

Course of Machine Learning
Master Degree in Computer Science
University of Rome ``Tor Vergata''

Giorgio Gambosi

a.a. 2017-2018

Model inference

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.

Model inference

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

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

Inference

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

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

Parameters

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 (Bayes)

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

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

where

- $p(\mathbf{Y})$ is the prior probability of $\mathbf{Y}$ (with respect to the observation of $\mathbf{X}$)
- $p(\mathbf{Y}|\mathbf{X})$ is the posterior probability of $\mathbf{Y}$
- $p(\mathbf{X}|\mathbf{Y})$ is the likelihood of $\mathbf{X}$ w.r.t. Y
- $p(\mathbf{X})$ is the evidence of $\mathbf{X}$

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

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

Maximum likelihood estimate

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 $\mathbf{x}$:

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

Maximum likelihood estimate

Example

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

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

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

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

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,

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

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

Example

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

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

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

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

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,

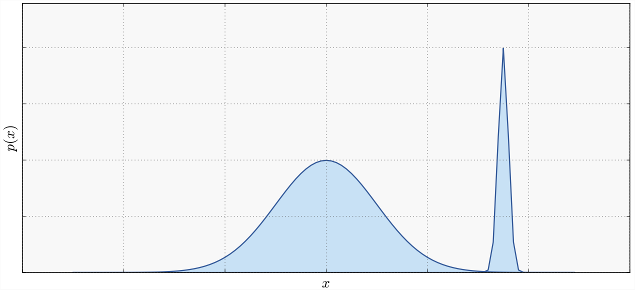
$$\Gamma(x) = (x-1)!$$

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

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

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

since $\int_{-\infty}^{+\infty} p(\phi|\mathbf{X},\alpha,\beta)d\phi = 1$, Z must be equal to the normalizing coefficient of the distribution $\text{Beta}(\phi|\alpha+N_1,\beta+N_0)$. Hence,

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

Model comparison

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

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

Example: learning in the dirichlet-multinomial model

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.

Language model

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

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.

Bayesian learning of a language model

We may apply the dirichlet-multinomial model:

- this implies defining a Dirichlet prior 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 Dir($\boldsymbol{\phi}|\boldsymbol{\alpha}'$), where

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

  that is,

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

Bayesian learning of a language model

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

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

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

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

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.

Naive bayes classifiers

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

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

By applying the product rule, it results

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

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 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 $\{I_1, \dots, I_n\}$

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

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.

Bayesian model comparison

Marginalization to reduce overfitting

- To avoid overfitting, we may apply marginalization of model parameters: this corresponds to averaging among all possible models
- Bayesian approach: use of probabilities to represent uncertainty in the choice of the model
- Set of L models $\mathcal{M}_i$, $i = 1, \ldots, L$, each a probability distribution over the observed data $\mathcal{T} = (\mathbf{X}, \mathbf{t})$ (conditional $p(\mathbf{t}|\mathbf{X})$ or joint $p(\mathbf{X}, \mathbf{t})$)
- Prior uncertainty about the model represented through distribution $p(\mathcal{M}_i)$
- Observing the training set modifies the uncertainty to the posterior

$$p(\mathcal{M}_i|\mathcal{T}) \propto p(\mathcal{T}|\mathcal{M}_i)p(\mathcal{M}_i)$$

- $p(\mathcal{T}|\mathcal{M}_i)$ is called marginal likelihood or model evidence
- $\dfrac{p(\mathcal{T}|\mathcal{M}_i)}{p(\mathcal{T}|\mathcal{M}_j)}$ is the Bayes factor for models $\mathcal{M}_i, \mathcal{M}_j$

Prediction

- Given the posterior among models, the predictive distribution can be obtained as a mixture distribution

$$p(t|\mathbf{x}, \mathcal{T}) = \sum_{i=1}^{L} p(t|\mathbf{x}, \mathcal{M}_i, \mathcal{T}) p(\mathcal{M}_i|\mathcal{T})$$

  this corresponds to a weighted average among predictions of single models, with weights given by their probabilities

As an average

- The evidence of a model can be expressed as an average among instances for all possible parameter values

$$p(\mathcal{T}|\mathcal{M}_i) = \int p(\mathcal{T}|\mathbf{w}, \mathcal{M}_i) p(\mathbf{w}|\mathcal{M}_i) d\mathbf{w}$$

  probability of generating $\mathcal{T}$ from a model with parameters derived by sampling distribution $p(\mathbf{w}|\mathcal{M}_i)$
- normalization term in definition of posterior distribution of parameters

$$p(\mathbf{w}|\mathcal{T}, \mathcal{M}_i) = \frac{p(\mathcal{T}|\mathbf{w}, \mathcal{M}_i) p(\mathbf{w}|\mathcal{M}_i)}{p(\mathcal{T}|\mathcal{M}_i)}$$

Insight

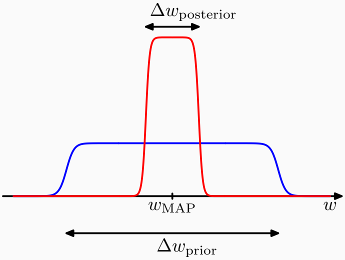
- Assume a model $\mathcal{M}$ with one parameter w
- Assume the posterior $p(w|\mathcal{T}, \mathcal{M}) \propto p(\mathcal{T}|w, \mathcal{M})p(w|\mathcal{M})$ is sharply peaked around w_{MAP}, with width Δw_{pos}, hence

$$\int p(\mathcal{T}|w, \mathcal{M})p(\mathbf{w}|\mathcal{M})dw \simeq p(\mathcal{T}|w_{MAP}, \mathcal{M})p(w_{MAP}|\mathcal{M})\Delta w_{pos}$$

- Assume also a flat prior $p(w|\mathcal{M})$ with width Δw_{pri} (and uniform probability $\frac{1}{\Delta w_{pri}}$) : then,

$$\begin{aligned} p(\mathcal{T}|\mathcal{M}) &= \int p(\mathcal{T}|w, \mathcal{M})p(w|\mathcal{M})dw \\ &\simeq p(\mathcal{T}|w_{MAP}, \mathcal{M})p(w_{MAP}|\mathcal{M})\Delta w_{pos} \simeq p(\mathcal{T}|w_{MAP}, \mathcal{M})\frac{\Delta w_{pos}}{\Delta w_{pri}} \end{aligned}$$

Model evidence



Model evidence

Taking logs,

$$\log p(\mathcal{T}|\mathcal{M}) \simeq \log p(\mathcal{T}|w_{MAP}, \mathcal{M}) + \log \frac{\Delta w_{pos}}{\Delta w_{pri}}$$

- The first term is the fit of data to the most probable parameter values
- The second term is negative ($\Delta w_{pos} < \Delta w_{pri}$) and it is a penalization related to the model complexity
    - Δw_{pos} very small: the parameter is finely tuned to data (even small differences in its value make the dataset unlikely). The second term is negative and large in module: the model is quite penalized
    - Δw_{pos} large: the parameter is only roughly tuned to data (the dataset has the same fit also for different parameter values). The second term is still negative, but small in module: the model has a small penalization

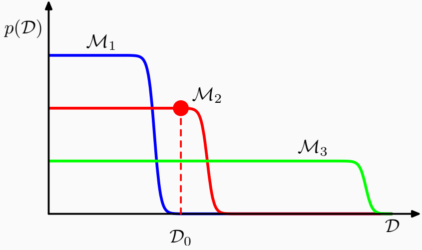
Model selection

Model with a set of M parameters, assuming all parameters have the same ratio $\frac{\Delta w_{pos}}{\Delta w_{pri}}$,

$$\log p(\mathcal{T}|\mathcal{M}) \simeq \log p(\mathcal{T}|\mathbf{w}_{MAP}, \mathcal{M}) + M \log \frac{\Delta w_{pos}}{\Delta w_{pri}}$$

Model complexities

- $\mathcal{M}_1$, low complexity: few datasets fitted ($\mathcal{D}_0$ does not fit)
- $\mathcal{M}_3$, high complexity: many datasets fitted, with low probability ($\mathcal{D}_0$ fits poorly)
- $\mathcal{M}_2$, intermediate complexity: some datasets fitted ($\mathcal{D}_0$ fits better)



Too simple model. If $\mathcal{M}_i$ is very simple, it will justify a limited collection of datasets (low generality) and $p(\mathcal{D}|\mathcal{M}_i)$ will assume significant values in a limited domain. Then, $p(\mathcal{D}_0|\mathcal{M}_i)$ will most likely be small, and $\mathcal{M}_i$ will not be selected.

Too complex model. If $\mathcal{M}_i$ is very complex, it will justify a large collection of datasets (high generality) and $p(\mathcal{D}|\mathcal{M}_i)$ will assume significant values in a large domain. As a consequence, such values will be small, since

$$\int_{\mathcal{D}} p(\mathcal{D}|\mathcal{M}_i) d\mathcal{D} = 1$$

Then, it is likely that $p(\mathcal{D}|\mathcal{M}_i)$ will be small, and $\mathcal{M}_i$ will not be selected.