Bayesian Models

Sandro Cumani

sandro.cumani@polito.it

Politecnico di Torino

Bayesian models

The frequentist approach to inference we have considered in previous lectures consists in estimating "optimal" model parameters by defining a suitable objective function and computing model parameters that optimize the objective function

A possible criterion consists in computing Maximum Likelihood estimates, i.e. estimating parameters that maximize the likelihood of our data (observed features in the generative framework, observed class labels in the discriminative framework)

ML point estimates, however, do not consider alternative, possible explanations of our data that may be less likely, but still properly model the distribution of our data — they neglect, as all point estimates, that our parameter estimates are uncertain

Bayesian models

The Bayesian framework allows taking into account our uncertainty over the model parameters

Since model parameters are not known, they are treated as uncertain quantities, that can be described in terms of probability distributions

Model parameters are thus represented in terms of Random Variables, and inference is based on probabilistic reasoning

Bayesian models

In the following we consider again a classification problem (although our discussion can be adapted to density estimation problems)

We also focus on generative approaches

We want to predict the class label for an unlabeled sample $\boldsymbol{x}_t$

We have at our disposal a labeled training set of samples

$$\mathcal{D} = \{(\boldsymbol{x}_1, c_1) \dots (\boldsymbol{x}_n, c_n)\}$$

or, in a more compact way:

$$\mathcal{D} = (\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{C}})$$

Bayesian models

For a generative model, we have seen that we can compute the class posterior distribution for the test samples from the joint distribution of observed features and class labels

Assuming that features are described by model $\mathcal{M}$ (i.e. a Gaussian model, a GMM model, ...), then inference is based on computing a predictive distribution

$$f_{\boldsymbol{X}_t, C_t|\boldsymbol{D},\mathcal{M}}(\boldsymbol{x}_t, c|\mathcal{D})$$

i.e., the joint distribution for features and classes, given the knowledge we have (the training set) and the model we chose ($\mathcal{M}$)

Bayesian models

The predictive distribution can be expressed as a ratio between the joint distribution of the observed samples (training samples and test samples) and labels (observed training labels, hypothesized test label c) and the joint distribution for the observed training data (Bayes rule):

$$f_{\boldsymbol{X}_t, C_t|\boldsymbol{D},\mathcal{M}}(\boldsymbol{x}_t, c|\boldsymbol{\mathcal{D}}) = \frac{f_{\boldsymbol{X}_t, C_t, \boldsymbol{D}|\mathcal{M}}(\boldsymbol{x}_t, c, \boldsymbol{\mathcal{D}})}{f_{\boldsymbol{D}|\mathcal{M}}(\boldsymbol{\mathcal{D}})}$$

Our problem thus consists in modeling the joint distributions of sets of feature vectors and corresponding labels

Bayesian models

Model $\mathcal{M}$ describes the joint distribution for a set of samples and classes in terms of a parametric distribution family

It allows defining conditional distributions of the form

$$f_{\boldsymbol{X}_1 \ldots \boldsymbol{X}_m, C_1 \ldots C_m | \boldsymbol{\Theta}}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_m, c_1, \ldots c_m | \boldsymbol{\theta})$$

where $\boldsymbol{X}_1 \ldots \boldsymbol{X}_m$ is a (generic) set of m R.V.s representing m features vectors, and $C_1 \ldots C_m$ are the R.V.s representing the corresponding labels

The model provides conditional distribution given a value $\boldsymbol{\theta}$ for the parameters

$\boldsymbol{\Theta}$ is the R.V. that describes the model parameters

Bayesian models

In the following we restrict the analysis to models that assume that observations are independent and identically distributed (i.i.d.) given the model parameters:

$$[(\boldsymbol{X}_i, C_i) \perp\!\!\!\perp (\boldsymbol{X}_j, C_j)] \,|(\boldsymbol{\Theta} = \boldsymbol{\theta})$$

and

$$(\boldsymbol{X}_i, C_i)|(\boldsymbol{\Theta} = \boldsymbol{\theta}) \sim (\boldsymbol{X}_j, C_j)|(\boldsymbol{\Theta} = \boldsymbol{\theta}) \sim (\boldsymbol{X}, C)|(\boldsymbol{\Theta} = \boldsymbol{\theta})$$

thus

$$f_{\boldsymbol{X}_1 \ldots \boldsymbol{X}_m, C_1 \ldots C_m|\boldsymbol{\Theta}}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_m, c_1, \ldots c_m|\boldsymbol{\theta}) = \prod_{i=1}^{m} f_{\boldsymbol{X}, C|\boldsymbol{\Theta}}(\boldsymbol{x}_i, c_i|\boldsymbol{\theta})$$

Bayesian models

Given a prior distribution $f_{\Theta}(\theta)$ for the model parameters Θ, we can then compute the joint distribution of a set of samples $\mathcal{D}_S = [\mathcal{X}_S, \mathcal{C}_S]$, with $\mathcal{X}_S = [\boldsymbol{x}_1 \dots \boldsymbol{x}_m]$, $\mathcal{C}_S = [c_1 \dots c_m]$:

$$\begin{aligned} f_{\boldsymbol{D}_S,\Theta}(\mathcal{D}_S, \theta) &= f_{\boldsymbol{X}_S,\boldsymbol{C}_S|\Theta}(\mathcal{X}_S, \mathcal{C}_S|\theta) f_{\Theta}(\theta) \\ &= \left[\prod_{i=1}^{m} f_{\boldsymbol{X}_i,C_i|\Theta}(\boldsymbol{x}_i, \boldsymbol{c}_i|\theta)\right] f_{\Theta}(\theta) \end{aligned}$$

The joint distribution consists of a product of sample likelihoods and parameters prior

The distribution for the data R.V. $\boldsymbol{D}_S$ is obtained by marginalizing with respect to the model parameters:

$$\begin{aligned} f_{\boldsymbol{D}_S}(\mathcal{D}_S) &= \int f_{\boldsymbol{D}_S,\Theta}(\mathcal{D}_S, \theta) d\theta \\ &= \int \left[\prod_{i=1}^{m} f_{\boldsymbol{X}_i,C_i|\Theta}(\boldsymbol{x}_i, \boldsymbol{c}_i|\theta)\right] f_{\Theta}(\theta) d\theta \end{aligned}$$

Bayesian models

We can apply these expressions to compute our predictive distribution (we drop the explicit conditioning on $\mathcal{M}$ for readability):

$$f_{\boldsymbol{X}_t, C_t|\boldsymbol{D}}(\boldsymbol{x}_t, c|\mathcal{D}) = \frac{f_{\boldsymbol{X}_t, C_t, \boldsymbol{D}}(\boldsymbol{x}_t, c, \mathcal{D})}{f_{\boldsymbol{D}}(\mathcal{D})}$$

with

$$f_{\boldsymbol{X}_t, C_t, \boldsymbol{D}}(\boldsymbol{x}_t, c, \mathcal{D}) = \int f_{\boldsymbol{X}_t, C_t|\boldsymbol{\Theta}}(\boldsymbol{x}_t, c_t|\boldsymbol{\theta}) f_{\boldsymbol{D}|\boldsymbol{\Theta}}(\mathcal{D}|\boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

$$f_{\boldsymbol{D}}(\mathcal{D}) = \int f_{\boldsymbol{D}|\boldsymbol{\Theta}}(\mathcal{D}|\boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

The factorizations derive from our independence assumptions

Bayesian models

We can also apply Bayes rule to write the predictive distribution in an equivalent form:

$$
\begin{aligned}
f_{\boldsymbol{X}_t, C_t|\boldsymbol{D}}(\boldsymbol{x}_t, c|\boldsymbol{\mathcal{D}}) &= \int f_{\boldsymbol{X}_t, C_t, \boldsymbol{\Theta}|\boldsymbol{D}}(\boldsymbol{x}_t, c, \boldsymbol{\theta}|\boldsymbol{\mathcal{D}}) d\boldsymbol{\theta} \\
&= \int f_{\boldsymbol{X}_t, C_t|\boldsymbol{\Theta}, \boldsymbol{D}}(\boldsymbol{x}_t, c|\boldsymbol{\theta}, \boldsymbol{\mathcal{D}}) f_{\boldsymbol{\Theta}|\boldsymbol{D}}(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}}) d\boldsymbol{\theta} \\
&= \int f_{\boldsymbol{X}_t, C_t|\boldsymbol{\Theta}}(\boldsymbol{x}_t, c|\boldsymbol{\theta}) f_{\boldsymbol{\Theta}|\boldsymbol{D}}(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}}) d\boldsymbol{\theta}
\end{aligned}
$$

The second step derives from the application of Bayes rule, whereas the third step derives from the assumption that samples are independent given the model parameters

The first term inside the integral is the likelihood

$$f_{\boldsymbol{X}_t, C_t|\boldsymbol{\Theta}}(\boldsymbol{x}_t, c|\boldsymbol{\theta})$$

If we assume that the prior class distribution is independent of the model parameters, we can also write

$$f_{\boldsymbol{X}_t, C_t|\boldsymbol{\Theta}}(\boldsymbol{x}_t, c|\boldsymbol{\theta}) = f_{\boldsymbol{X}_t|C_t,\boldsymbol{\Theta}}(\boldsymbol{x}_t|c, \boldsymbol{\theta})P(C_t = c)$$

i.e., the product of class-conditional likelihoods and class priors.

The predictive distribution becomes

$$\int f_{\boldsymbol{X}_t|C_t,\boldsymbol{\Theta}}(\boldsymbol{x}_t|c, \boldsymbol{\theta}) f_{\boldsymbol{\Theta}|\boldsymbol{D}}(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})d\boldsymbol{\theta} \cdot P(C_t = c)$$

i.e. a product of the class prior and the class-conditional predictive distribution

$$f_{\boldsymbol{X}_t|C_t,\boldsymbol{D}}(\boldsymbol{x}_t|c, \boldsymbol{\mathcal{D}}) = \int f_{\boldsymbol{X}_t|C_t,\boldsymbol{\Theta}}(\boldsymbol{x}_t|c, \boldsymbol{\theta}) f_{\boldsymbol{\Theta}|\boldsymbol{D}}(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})d\boldsymbol{\theta}$$

Bayesian models

The second term

$$f_{\boldsymbol{\Theta}|\boldsymbol{D}}(\boldsymbol{\theta}|\mathcal{D})$$

is the posterior distribution for the model parameters given the training data (and the chosen model $\mathcal{M}$)

It represents our belief over the possible model parameters, given the training dataset

Before we observe the training data, our belief corresponds to the prior distribution $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$

Once we have observed data $\boldsymbol{D} = \mathcal{D}$, our belief is updated to $f_{\boldsymbol{\Theta}|\boldsymbol{D}}(\boldsymbol{\theta}|\mathcal{D})$

Bayesian models

The posterior distribution can be expressed in terms of conditional likelihoods and priors:

$$f_{\Theta|\boldsymbol{D}}(\theta|\mathcal{D}) = \frac{f_{\Theta,\boldsymbol{D}}(\theta,\mathcal{D})}{f_{\boldsymbol{D}}(\mathcal{D})} = \frac{f_{\boldsymbol{D}|\Theta}(\mathcal{D}|\theta)f_{\Theta}(\theta)}{\int f_{\boldsymbol{D}|\Theta}(\mathcal{D}|\theta)f_{\Theta}(\theta)d\theta}$$
$$\propto f_{\boldsymbol{D}|\Theta}(\mathcal{D}|\theta)f_{\Theta}(\theta)$$

The posterior distribution is proportional to the product of the likelihood term

$$f_{\boldsymbol{D}|\Theta}(\mathcal{D}|\theta)$$

and the prior distribution

$$f_{\Theta}(\theta)$$

The proportionality factor is the value that ensures that the posterior distribution integrates to 1.

Bayesian models

In the Bayesian framework "learning" consists in updating our beliefs as we collect more evidence

For example, if we have computed the posterior distribution given a set $\mathcal{D}_1$ of training data, and we then collect additional data $\mathcal{D}_2$, we can refine our beliefs as

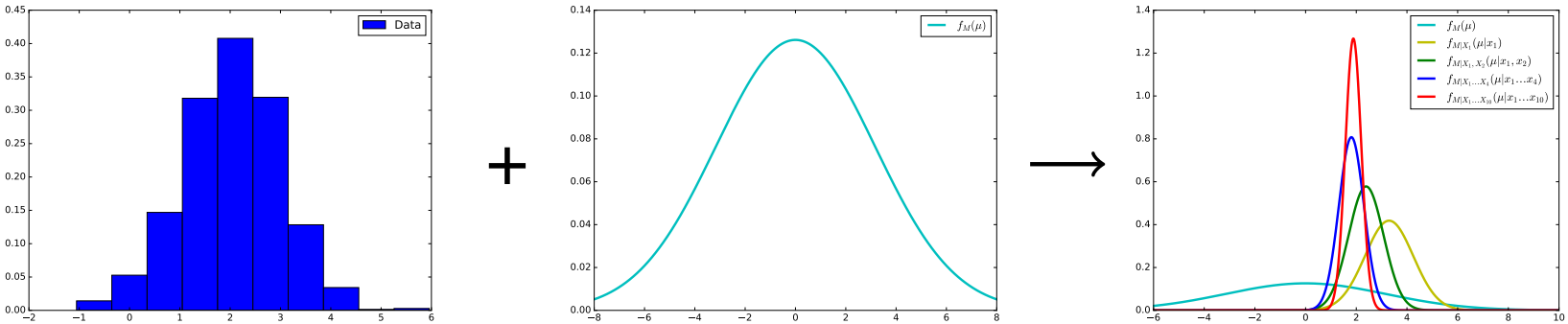
$$\begin{aligned} f_{\Theta|\boldsymbol{D}_1,\boldsymbol{D}_2}(\theta|\mathcal{D}_1,\mathcal{D}_2) &\propto f_{\boldsymbol{D}_1|\Theta}(\mathcal{D}_1|\theta) f_{\boldsymbol{D}_2|\Theta}(\mathcal{D}_2|\theta) f_{\Theta}(\theta) \\ &\propto f_{\boldsymbol{D}_2|\Theta}(\mathcal{D}_2|\theta) f_{\Theta|\boldsymbol{D}_1}(\theta|\mathcal{D}_1) \end{aligned}$$

i.e., the posterior given the two datasets is proportional to the product of likelihood computed over the new data and the posterior distribution given the initial dataset

The posterior distribution given dataset $\boldsymbol{D}_1 = \mathcal{D}_1$ acts as a prior for the remaining data, and collects all the information we have extracted from the initial dataset

Bayesian models

From prior to posterior distributions — modeling the mean of a Gaussian distribution:



We can observe that, as we increase the data size, the posterior distribution tends to concentrate more and more around a single value

This corresponds to a decrease in our uncertainty over the possible model parameters

Bayesian models

In theory, once we have chosen a prior distribution, we have all the elements to compute the predictive distribution and thus to perform inference

However, in practice computing predictive distributions requires solving integrals, which may prove a difficult task

Furthermore, the results will depend on our choice of the prior, thus we need to decide what kind of prior is appropriate for our task

Bayesian models

The first issue can be addressed by using conjugate prior distributions

We say that a prior distribution is conjugate to a likelihood function if the corresponding posterior distribution belongs to the same family as the prior distribution, i.e. they have the same functional form

In this case, since we know the posterior distribution up to a normalization factor, the normalization factor can be recovered from the distribution family

Bayesian models

For example, we can consider a Gaussian likelihood for a (unlabeled) dataset $\boldsymbol{\mathcal{X}}$, parametrized by the distribution mean (we assume that the covariance matrix is known and fixed)

Let the log-likelihood (a function of the mean $\boldsymbol{\mu}$) be

$$\begin{aligned}\ell(\boldsymbol{\mu}) &= \log \mathcal{L}(\boldsymbol{\mu}) = \log f_{\boldsymbol{X}|\boldsymbol{M}}(\boldsymbol{\mathcal{X}}|\boldsymbol{\mu}) \\ &= \sum_i \frac{1}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Lambda}\,(\boldsymbol{x}_i - \boldsymbol{\mu}) + \xi_1 \\ &= \boldsymbol{\mu}^T \boldsymbol{\Lambda} \boldsymbol{F} - \frac{n}{2}\boldsymbol{\mu}^T \boldsymbol{\Lambda}\boldsymbol{\mu} + \xi_2\end{aligned}$$

where $F = \sum_i \boldsymbol{x}_i$, n is the number of samples in $\boldsymbol{\mathcal{X}}$ and ξ_2 collects all terms that do not depend on $\boldsymbol{\mu}$

Bayesian models

The posterior distribution for $\boldsymbol{\mu}$ is proportional to the product of the prior distribution and the likelihood.

Let the prior distribution belong to the Gaussian family

$$f_{\boldsymbol{M}}(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{m}, \boldsymbol{P}^{-1})$$

where $\boldsymbol{m}$ and $\boldsymbol{P}$ are the mean and precision matrix of the prior

The functional form of the prior distribution is

$$\log f_{\boldsymbol{M}}(\boldsymbol{\mu}) = \boldsymbol{\mu}^T \boldsymbol{P}\boldsymbol{m} - \frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{P}\boldsymbol{\mu} + \xi$$

A log-density defined, up to a constant term, by the above expression corresponds to a normal density $\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{m}, \boldsymbol{P}^{-1})$

The constant term ξ corresponds to the logarithm of the normalization constant

The posterior distribution for $\boldsymbol{M}$ corresponds to

$$f_{\boldsymbol{M}|\boldsymbol{X}}(\boldsymbol{\mu}|\boldsymbol{\mathcal{X}}) \propto \mathcal{L}(\boldsymbol{\mu})\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{m}, \boldsymbol{P}^{-1})$$

In log form

$$\begin{aligned}\log f_{\boldsymbol{M}|\boldsymbol{X}}(\boldsymbol{\mu}|\boldsymbol{\mathcal{X}}) &= \ell(\boldsymbol{\mu}) - \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{P}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{P}\boldsymbol{m} + \xi_3 \\ &= \boldsymbol{\mu}^T\left(\boldsymbol{\Lambda}\boldsymbol{F} + \boldsymbol{P}\boldsymbol{m}\right) - \frac{1}{2}\boldsymbol{\mu}^T\left(n\boldsymbol{\Lambda} + \boldsymbol{P}\right)\boldsymbol{\mu} + \xi_4\end{aligned}$$

The expression has the functional form of a multivariate Gaussian distribution:

$$f_{\boldsymbol{M}|\boldsymbol{X}}(\boldsymbol{\mu}|\boldsymbol{\mathcal{X}}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{m}_{\boldsymbol{x}}, \boldsymbol{P}_{\boldsymbol{x}}^{-1}) \iff \log f_{\boldsymbol{M}|\boldsymbol{X}}(\boldsymbol{\mu}|\boldsymbol{\mathcal{X}}) = \boldsymbol{\mu}^T\boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{m}_{\boldsymbol{x}} - \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{P}_{\boldsymbol{x}}\boldsymbol{\mu} + \xi$$

The posterior distribution for the mean is therefore again a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{m_x}, \boldsymbol{P_x}^{-1})$

The parameters can be recovered by inspection:

$$\boldsymbol{P_x} = \boldsymbol{P} + n\boldsymbol{\Lambda}$$
$$\boldsymbol{m_x} = \boldsymbol{P_x}^{-1}(\boldsymbol{\Lambda F} + \boldsymbol{Pm}) = (\boldsymbol{P} + n\boldsymbol{\Lambda})^{-1}(\boldsymbol{\Lambda F} + \boldsymbol{Pm})$$

Using conjugate priors we can avoid solving difficult integrals — the posterior is in the same family as the prior, so we already know the values for the normalization constant

With conjugate priors we are also able to compute the integral required for the predictive distribution

We can observe that we can also express the predictive distribution in terms of posterior distributions using Bayes rule:

$$f_{X_t|C_t,\boldsymbol{D}}(\boldsymbol{x}_t|c,\boldsymbol{\mathcal{D}}) = \frac{f_{X_t|C_t,\boldsymbol{\Theta}}(\boldsymbol{x}_t|c,\boldsymbol{\theta}) f_{\boldsymbol{\Theta}|\boldsymbol{D}}(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})}{f_{\boldsymbol{\Theta}|\boldsymbol{D},X_t,C_t}(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}},\boldsymbol{x}_t,c)}$$

The expression is valid for any value of $\boldsymbol{\theta}$ for which the denominator is non-zero, and provides an alternative to solving the predictive distribution integral (in practice, we already solved the integration when computing the posterior density)

Let's see what we get if we use the ML solution for θ to compute the previous expression:

$$f_{X_t|C_t,\boldsymbol{D}}(\boldsymbol{x}_t|c,\boldsymbol{\mathcal{D}}) = f_{X_t|C_t,\boldsymbol{\Theta}}(\boldsymbol{x}_t|c,\boldsymbol{\theta}^*)\frac{f_{\boldsymbol{\Theta}|\boldsymbol{D}}(\boldsymbol{\theta}^*|\boldsymbol{\mathcal{D}})}{f_{\boldsymbol{\Theta}|\boldsymbol{D},X_t,C_t}(\boldsymbol{\theta}^*|\boldsymbol{\mathcal{D}},\boldsymbol{x}_t,c)}$$

We can contrast with the ML framework: our predictive distribution was

$$f_{X_t|C_t,\boldsymbol{\Theta}}(\boldsymbol{x}_t|c,\boldsymbol{\theta}^*)$$

For a binary classification problem we can see that the log-likelihood would be

$$\log\frac{f_{\boldsymbol{x}_t|C_t,\boldsymbol{D}}(\boldsymbol{x}_t|\mathcal{H}_T,\boldsymbol{\mathcal{D}})}{f_{\boldsymbol{x}_t|C_t,\boldsymbol{D}}(\boldsymbol{x}_t|\mathcal{H}_F,\boldsymbol{\mathcal{D}})} = \log\frac{f_{X_t|C_t,\boldsymbol{\Theta}}(\boldsymbol{x}_t|\mathcal{H}_T,\boldsymbol{\theta}^*)}{f_{X_t|C_t,\boldsymbol{\Theta}}(\boldsymbol{x}_t|\mathcal{H}_F,\boldsymbol{\theta}^*)} - \log\frac{f_{\boldsymbol{\Theta}|\boldsymbol{D},X_t,C_t}(\boldsymbol{\theta}^*|\boldsymbol{\mathcal{D}},\boldsymbol{x}_t,\mathcal{H}_T)}{f_{\boldsymbol{\Theta}|\boldsymbol{D},X_t,C_t}(\boldsymbol{\theta}^*|\boldsymbol{\mathcal{D}},\boldsymbol{x}_t,\mathcal{H}_F)}$$

Bayesian models

The prior distribution encodes our prior beliefs over the model parameters, and can have a large impact over the model when we have limited amounts of data

Priors may be chosen to encode expert knowledge

In other cases, we may not have strong indications on preferred models

What kind of priors should we use?

Bayesian models

If we have no reason to choose a strongly informative prior, we may resort to non-informative or weakly informative priors

- Flat prior $f_{\Theta}(\boldsymbol{\theta}) \propto \alpha$
- Jeffrey's priors
- ...

The flat prior, for example, encodes that all values of the parameters are, a priori, equally likely

We should note that a flat prior is often improper, in that it cannot be normalized, since its integral diverges (e.g. this happens if we use a flat prior over $\mathbb{R}$)

Improper priors may still, in many cases, provide proper posteriors, and thus may be still used in the Bayesian framework

In practice, it may be better to represent an improper prior as a limit of a proper prior. For example, we may represent an "almost" flat prior as

$$f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \mathcal{N}(0, \varepsilon^{-1}\boldsymbol{I})$$

and then let $\varepsilon \to 0$

In this case, it's better to work with the proper prior and pass to the limit only at the end

Bayesian models

Another option is to use empirical priors: we can represent the prior as a parametric distribution, and learn the parameters of the prior distribution from the data, for example maximizing the marginal likelihood

These kind of models are also known as empirical Bayes approaches

We will see an example shortly

Bayesian models

The frequentist approach we have discussed in previous lectures can be interpreted as an approximation of the Bayesian approach

We can interpret frequentist estimators as approximations of posterior distribution

$$f_{\boldsymbol{\Theta}|\boldsymbol{D}}(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}}) \approx \delta_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})$$

where the posterior is approximated by a delta distribution centered around the estimated value $\boldsymbol{\theta}^*$

The predictive distribution becomes

$$f_{\boldsymbol{X}_t|C_t,\boldsymbol{D}}(\boldsymbol{x}_t|c,\boldsymbol{\mathcal{D}}) = \int f_{\boldsymbol{X}_t|C_t,\boldsymbol{\Theta}}(\boldsymbol{x}_t|c,\boldsymbol{\theta})\delta_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})d\boldsymbol{\theta} = f_{\boldsymbol{X}_t|C_t,\boldsymbol{\Theta}}(\boldsymbol{x}_t|c,\boldsymbol{\theta}^*)$$

Bayesian models

Since the delta distribution is approximating our posterior, we may select as value for θ^* the value that is most likely according to the posterior distribution

$$\begin{aligned}\boldsymbol{\theta}^*_{MAP} &= \arg\max_{\boldsymbol{\theta}} f_{\boldsymbol{\Theta}|\boldsymbol{D}}(\boldsymbol{\theta}|\mathcal{D}) \\ &= \arg\max_{\boldsymbol{\theta}} f_{\boldsymbol{D}|\boldsymbol{\Theta}}(\mathcal{D}|\boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\end{aligned}$$

The estimator θ^*_{MAP} is called Maximum A Posteriori (MAP) estimator, since its the most likely value according to the posterior distribution

The approximation will become more precise as the number of samples increases, since the variance of the posterior distribution will tend to decrease

Bayesian models

If we choose a flat prior $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \propto \alpha$, then the MAP estimate becomes

$$\boldsymbol{\theta}^*_{ML} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$$

i.e., the ML estimator that maximizes the likelihood of the training data

We can thus re-interpret ML estimators as an approximation of the Bayesian posterior assuming flat priors

The approximation is good if we have a large number of training samples

Bayesian models

We can also re-interpret the regularized logistic regression model

$$\arg\min_{\boldsymbol{w},b} \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \sum_i \log\left[1 + e^{-z_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right)}\right]$$

as a MAP estimate, where the prior for the model parameters $\boldsymbol{w}$ is a Gaussian distribution

$$f_{\boldsymbol{W}}(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

Bayesian models

We now consider some examples that allow us to better understand how to perform inference using Bayesian models

We start considering again the multinomial model for text classification

The multinomial model assumes that tokens of each class can be modeled by multinomial distributions with parameters $\boldsymbol{\pi}_1 \dots \boldsymbol{\pi}_K$, where K is the number of classes

The log-likelihood of the model is

$$\ell(\boldsymbol{\pi}_1, \dots \boldsymbol{\pi}_K) = \sum_{c=1}^{K} \sum_{j=1}^{m} N_{c,j} \log \pi_{c,j} + \xi$$

Bayesian models

In the following, we assume that the model parameters are represented by i.i.d. R.V.s $\boldsymbol{\Pi}_1 \dots \boldsymbol{\Pi}_K \sim \boldsymbol{\Pi}$, with prior distribution given by

$$f_{\boldsymbol{\Pi}_i}(\boldsymbol{\pi}) = f_{\boldsymbol{\Pi}}(\boldsymbol{\pi})$$

We can show that a conjugate prior for $\boldsymbol{\Pi}_1 \dots \boldsymbol{\Pi}_K$ is the Dirichlet prior

$$\boldsymbol{\Pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}) \ , \quad f_{\boldsymbol{\Pi}}(\boldsymbol{\pi}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{m} \boldsymbol{\pi}_i^{\boldsymbol{\alpha}_i - 1}$$

where $B(\boldsymbol{\alpha})$ is the Beta function

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{m} \Gamma(\boldsymbol{\alpha}_i)}{\Gamma\left(\sum_{i=1}^{m} \boldsymbol{\alpha}_i\right)}$$

Γ is the Gamma function, a generalization of the factorial function to non-integer values — for example, $\Gamma(n) = (n-1)!$ and $\Gamma(x) = x\Gamma(x-1)$

$\boldsymbol{\alpha}$ is a vector of m positive, real values that parametrize the Dirichlet density, and m is the number of unique tokens (size of the dictionary)

The prior log-density for $\mathbf{\Pi}_c$ therefore has the form

$$\log f_{\mathbf{\Pi}_c|\boldsymbol{\alpha}}(\boldsymbol{\pi}_c|\boldsymbol{\alpha}) = \sum_{i=1}^{m}(\boldsymbol{\alpha}_i - 1)\log \boldsymbol{\pi}_{c,i} + \xi$$

for some constant[1] ξ

Since we assume that $\mathbf{\Pi}_1 \dots \mathbf{\Pi}_K$ are i.i.d., the prior log-density is

$$\log f_{\mathbf{\Pi}_1 \dots \mathbf{\Pi}_K|\boldsymbol{\alpha}}(\boldsymbol{\pi}_1 \dots \boldsymbol{\pi}_K|\boldsymbol{\alpha}) = \sum_{c=1}^{K}\sum_{i=1}^{m}(\boldsymbol{\alpha}_i - 1)\log \boldsymbol{\pi}_{c,i} + \xi$$

[1] In the following we will denote all irrelevant constants with ξ, even if they refer to different actual quantities

The posterior distribution is proportional to the product of likelihood and prior, thus

$$\begin{aligned}
&\log f_{\boldsymbol{\Pi}_1 \ldots \boldsymbol{\Pi}_K | \boldsymbol{D}, \boldsymbol{\alpha}}(\boldsymbol{\pi}_1 \ldots \boldsymbol{\pi}_K | \mathcal{D}, \boldsymbol{\alpha}) \\
&\quad = \sum_{c=1}^{K} \sum_{i=1}^{m} N_{c,i} \log \pi_{c,i} + \sum_{c=1}^{K} \sum_{i=1}^{m} (\boldsymbol{\alpha}_i - 1) \log \boldsymbol{\pi}_{c,i} + \xi \\
&\quad = \sum_{c=1}^{K} \sum_{i=1}^{m} (N_{c,i} + \alpha_i - 1) \log \pi_{c,j} + \xi
\end{aligned}$$

We can observe that the posterior distribution factorizes over the terms $\boldsymbol{\Pi}_c$ of the different classes:

$$\log f_{\boldsymbol{\Pi}_1 \ldots \boldsymbol{\Pi}_K | \boldsymbol{D}, \boldsymbol{\alpha}}(\boldsymbol{\pi}_1 \ldots \boldsymbol{\pi}_K | \mathcal{D}, \boldsymbol{\alpha}) = \sum_{c=1}^{K} \log f_{\boldsymbol{\Pi}_c | \boldsymbol{D}_c, \boldsymbol{\alpha}}(\boldsymbol{\pi}_c | \mathcal{D}_c, \boldsymbol{\alpha})$$

where $\mathcal{D}_c$ are the training samples for class c

Bayesian models

The terms $f_{\boldsymbol{\Pi}_c|\boldsymbol{D}_c,\boldsymbol{\alpha}}(\boldsymbol{\pi}_c|\boldsymbol{\mathcal{D}}_c,\boldsymbol{\alpha})$ are the posterior distributions of the different model parameters, given the data of the corresponding class

$$f_{\boldsymbol{\Pi}_c|\boldsymbol{D}_c,\boldsymbol{\alpha}}(\boldsymbol{\pi}_c|\boldsymbol{\mathcal{D}}_c,\boldsymbol{\alpha}) = \sum_{i=1}^{m}(N_{c,i}+\alpha_i-1)\log\pi_{c,i}+\xi$$

We can recognize that this is the log-density of a Dirichlet distribution

$$\boldsymbol{\Pi}_c|\boldsymbol{D}_c = \boldsymbol{\mathcal{D}}_c,\boldsymbol{\alpha} \sim \mathrm{Dir}(\boldsymbol{\alpha}+\boldsymbol{N}_c)$$

where $\boldsymbol{N}_c = [N_{c,1}\ldots N_{c,m}]$

We can interpret the parameters of the posterior distribution as counts of each token. The term $\boldsymbol{\alpha}$ represents "pseudo-counts" that we are adding to the counts

Before we consider the predictive distribution, we focus on the selection of the values $\boldsymbol{\alpha}$ of the prior distribution

From the form of the density of the Dirichlet distribution, we can observe that setting

$$\boldsymbol{\alpha}_i = 1 \ , \quad \forall i$$

corresponds to a uniform (flat) prior

$$f_{\boldsymbol{\Pi}|\boldsymbol{\alpha}=\mathbf{1}}(\boldsymbol{\pi}|\boldsymbol{\alpha} = \mathbf{1}) = \frac{1}{B(\mathbf{1})}$$

In this case, the prior is proper, since the support for $\boldsymbol{\Pi}$ is the unit-box

If we choose to use the flat prior, then the posterior distribution becomes

$$\boldsymbol{\Pi}_c | \boldsymbol{D}_c = \boldsymbol{\mathcal{D}}_c, \boldsymbol{\alpha} = \mathbf{1} \sim \mathrm{Dir}(\boldsymbol{N}_c + \mathbf{1})$$

The MAP solution in this case corresponds to the ML solution

However, we can observe that the mean of the posterior distribution is a vector with components

$$\mathbb{E}\left[\boldsymbol{\Pi}_c | \boldsymbol{D}_c = \boldsymbol{\mathcal{D}}_c, \boldsymbol{\alpha} = \mathbf{1}\right]_i = \frac{\boldsymbol{N}_{c,i} + 1}{\sum_{j=1}^{m} \boldsymbol{N}_{c,j} + 1}$$

If, rather than the MAP / ML estimates, we approximate the posterior with the posterior mean, then we retrieve the pseudo-count solution with pseudo-counts $\varepsilon = 1.0$ of Laboratory 6

The same solution corresponds also to a MAP estimate, but with a non-flat prior $\boldsymbol{\alpha}_i = 1 + \varepsilon$

The predictive distribution can be obtained by integrating the product of the class-conditional likelihood for the test sample and the posterior probability of the parameters given the training samples.

Let $\boldsymbol{y}_t$ be the vector of token occurrences for the test document. The predictive distribution is

$$f_{\boldsymbol{Y}_t|C,\boldsymbol{D},\boldsymbol{\alpha}}(\boldsymbol{y}_t|c,\mathcal{D},\boldsymbol{\alpha}) = \int f_{\boldsymbol{Y}_t|\boldsymbol{\Pi}_c}(\boldsymbol{y}_t|\boldsymbol{\pi}) f_{\boldsymbol{\Pi}_c|\boldsymbol{D}_c,\boldsymbol{\alpha}}(\boldsymbol{\pi}|\mathcal{D}_c,\boldsymbol{\alpha}) d\boldsymbol{\pi}$$

The first term in the integral is a multinomial p.d.f., whereas the second term is the Dirichlet density of the posterior distribution for the parameters. We can show that

$$f_{\boldsymbol{Y}_t|C,\boldsymbol{D},\boldsymbol{\alpha}}(\boldsymbol{y}_t|c,\mathcal{D},\boldsymbol{\alpha}) = \frac{\left(\sum_{j=1}^{m} \boldsymbol{y}_{t,j}\right) B\left(\sum_{j=1}^{m}(N_{c,j}+\boldsymbol{\alpha}_j), \sum_{j=1}^{m} \boldsymbol{y}_{t,j}\right)}{\prod_{j=1}^{m} \boldsymbol{y}_{t,j} B\left(N_{c,j}+\boldsymbol{\alpha}_j, \boldsymbol{y}_{t,j}\right)}$$

Bayesian models

Compared to the ML solution, the Bayesian approach can provide better results. For the task of Laboratory 6 we have

Model	Accuracy
ML $\varepsilon = 10^{-6}$	52%
Mean point estimate (ML + pseudo-counts $\varepsilon = 1$)	58%
Bayes, flat prior $\boldsymbol{\alpha}_i = 1$	58%

The Bayesian approach achieves better solution than the ML solution (note: the ML solution is actually already using small pseudo-counts to avoid numerical issues)

The posterior mean provides a good approximation of the posterior distribution, and indeed allows us to obtain the same results as for the Bayesian model

Bayesian models

The second example we analyze considers a tied covariance Gaussian model, but introduces a Bayesian modeling of the class means

As we have seen, the tied covariance model can be interpreted as an extension of LDA

The model we are going to consider can thus be interpreted as a probabilistic version of LDA, in the sense that class means will be modeled in a probabilistic way

As we shall see, Bayesian treatment of the class means will also allow us to extend the model to open-set classification

Bayesian models

We consider a K-class problem, where the samples of each class are modeled by a Gaussian density

$$f_{\boldsymbol{X}_i|C,\boldsymbol{\mu}_1\ldots\boldsymbol{\mu}_K,\boldsymbol{\Lambda}}(\boldsymbol{x}_i|c,\boldsymbol{\mu}_1\ldots\boldsymbol{\mu}_K,\boldsymbol{\Lambda}) = \mathcal{N}\left(\boldsymbol{x}_i|\boldsymbol{\mu}_c,\boldsymbol{\Lambda}^{-1}\right)$$

We also adopt a slightly different notation, and represent with $\boldsymbol{x}_{c,i}$ the i-th training set sample of class c

We have seen that the model allows interpreting a sample as the sum of the corresponding class mean and i.i.d., zero-mean, Gaussian distributed noise

$$\boldsymbol{x}_{c,i} = \boldsymbol{\mu}_c + \boldsymbol{\varepsilon}_{c,i}\ , \quad \boldsymbol{\varepsilon}_{c,i} \sim \boldsymbol{E}_{c,i} \sim \mathcal{N}\left(\boldsymbol{0},\boldsymbol{\Lambda}^{-1}\right)$$

$\boldsymbol{E}_{c,i}$ is the R.V. that represents the noise. $\boldsymbol{\varepsilon}_{c,i}$ is the noise realization

Bayesian models

The frequentist approach we have followed consisted in performing ML estimation of the class means $\boldsymbol{M}$ and of the noise covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1}$

We now consider a Bayesian treatment of the class means

Again, we resort to a conjugate prior to be able to compute closed form solutions for our integrals

We have seen that a Gaussian distribution is a conjugate prior for the mean of a Gaussian likelihood

We can use a parametric prior, and assume that the means are a priori distributed as

$$\boldsymbol{M}_1 \sim \boldsymbol{M}_2 \sim \cdots \sim \boldsymbol{M}_K \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{B})$$

Our model is thus specified by the likelihood function

$$f_{\boldsymbol{X}|C,\boldsymbol{M}_1\ldots\boldsymbol{M}_K}(\boldsymbol{x}|c,\boldsymbol{\mu}_1\ldots\boldsymbol{\mu}_K)=\mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_c,\boldsymbol{\Lambda}^{-1}\right)$$

and the prior for the class means

$$f_{\boldsymbol{M}_c|\boldsymbol{m},\boldsymbol{B}}(\boldsymbol{\mu}|\boldsymbol{m},\boldsymbol{B})=\mathcal{N}(\boldsymbol{m},\boldsymbol{B})$$

The complete data likelihood for the training set is given by

$$f_{\boldsymbol{D},\boldsymbol{M}_1\ldots\boldsymbol{M}_K|\boldsymbol{m},\boldsymbol{B},\boldsymbol{\Lambda}}(\mathcal{D},\boldsymbol{\mu}_1\ldots\boldsymbol{\mu}_K|\boldsymbol{m},\boldsymbol{B},\boldsymbol{\Lambda})=$$
$$\prod_{c=1}^{K}\left[\left(\prod_{i=1}^{n_c}\mathcal{N}(\boldsymbol{x}_{c,i}|\boldsymbol{\mu}_c,\boldsymbol{\Lambda}^{-1})\right)\mathcal{N}(\boldsymbol{\mu}_c|\boldsymbol{m},\boldsymbol{P})\right]$$

We can see that the likelihood factorizes over classes, and thus the same applies to the posterior distribution for $\boldsymbol{M}_1\ldots\boldsymbol{M}_K$

Bayesian models

For each term $\boldsymbol{M}_i$, the complete data likelihood is proportional to a product of a Gaussian likelihood and a Gaussian prior

We have already shown that this corresponds to a Gaussian posterior distribution:

$$f_{\boldsymbol{M}_c|\boldsymbol{D},\boldsymbol{m},\boldsymbol{P},\boldsymbol{\Lambda}}(\boldsymbol{\mu}_c|\mathcal{D},\boldsymbol{m},\boldsymbol{P},\boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{\mu}_c|\boldsymbol{m}_c,\boldsymbol{P}_c^{-1})$$

with

$$\boldsymbol{P}_c = \boldsymbol{P} + n_c\boldsymbol{\Lambda}$$
$$\boldsymbol{m}_c = \boldsymbol{P}_c^{-1}\left(\boldsymbol{\Lambda}\sum_i \boldsymbol{x}_{c,i} + \boldsymbol{P}\boldsymbol{m}\right)$$

The class-conditional predictive distribution can be computed in closed form as

$$
\begin{aligned}
& f_{\boldsymbol{X}_t|C_t,\boldsymbol{D},\boldsymbol{m},\boldsymbol{B},\boldsymbol{\Lambda}}(\boldsymbol{x}_t|c,\mathcal{D},\boldsymbol{m},\boldsymbol{B},\boldsymbol{\Lambda}) \\
&\quad = \int f_{\boldsymbol{X}_t|C_t,\boldsymbol{M}_c,\boldsymbol{\Lambda}}(\boldsymbol{x}_t|c,\boldsymbol{\mu},\boldsymbol{\Lambda}) f_{\boldsymbol{M}_c|\boldsymbol{D},\boldsymbol{m},\boldsymbol{B},\boldsymbol{\Lambda}}(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\mu},\boldsymbol{B},\boldsymbol{\Lambda}) d\boldsymbol{\mu} \\
&\quad = \int \mathcal{N}\left(\boldsymbol{x}_t|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}\right) \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{m}_c,\boldsymbol{P}_c^{-1}\right) d\boldsymbol{\mu} \\
&\quad = \mathcal{N}\left(\boldsymbol{x}_t|\boldsymbol{m}_c,\boldsymbol{\Lambda}^{-1}+\boldsymbol{P}_c^{-1}\right)
\end{aligned}
$$

Bayesian models

The model we derived depends on the values of the likelihood parameter $\boldsymbol{\Lambda}$ and the prior parameters $\boldsymbol{m}, \boldsymbol{P}$

A Bayesian treatment of $\boldsymbol{\Lambda}$ would make the problem much harder, therefore we adopt, in this example, a frequentist approach for the estimation of $\boldsymbol{\Lambda}$

We still have to specify the prior parameters $\boldsymbol{m}, \boldsymbol{P}$

Rather than fixing the values of $\boldsymbol{m}$ and $\boldsymbol{P}$, we follow an empirical Bayes approach and we try to estimate the model parameters from the data (an alternative would be to use improper priors, or use hyper-priors to model also the parameters $\boldsymbol{m}$ and $\boldsymbol{P}$ is a Bayesian way)

We consider again Maximum Likelihood estimates

We compute Λ, $\boldsymbol{m}$ and $\boldsymbol{P}$ as the maximizer of the likelihood of the observed data, i.e. after we have marginalized over the latent factors $\boldsymbol{M}_c$

$$\begin{aligned}\mathcal{L}_{\boldsymbol{D}}(\boldsymbol{m},\boldsymbol{P},\Lambda) &= f_{\boldsymbol{D}|\boldsymbol{m},\boldsymbol{P},\Lambda}(\mathcal{D}|\boldsymbol{m},\boldsymbol{B},\Lambda) \\ &= \prod_{c=1}^{K}\int\prod_{i=1}^{n_c}\mathcal{N}\left(\boldsymbol{x}_{c,i}|\boldsymbol{\mu},\Lambda^{-1}\right)\mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{m},\boldsymbol{P}^{-1}\right)d\boldsymbol{\mu}\end{aligned}$$

Direct maximization of the likelihood is not straightforward. However, the complete data likelihood has a simple expression

$$\begin{aligned}\mathcal{L}_{\boldsymbol{D},\boldsymbol{M}}(\boldsymbol{m},\boldsymbol{P},\Lambda) &= f_{\boldsymbol{D},\boldsymbol{M}|\boldsymbol{m},\boldsymbol{P},\Lambda}(\mathcal{D},\boldsymbol{\mu}_1\ldots\boldsymbol{\mu}_K|\boldsymbol{m},\boldsymbol{P},\Lambda) \\ &= \prod_{c=1}^{K}\left[\left(\prod_{i=1}^{n_c}\mathcal{N}\left(\boldsymbol{x}_{c,i}|\boldsymbol{\mu}_c,\Lambda^{-1}\right)\right)\mathcal{N}\left(\boldsymbol{\mu}_c|\boldsymbol{m},\boldsymbol{P}^{-1}\right)\right]\end{aligned}$$

In log-form:

$$\ell_{\boldsymbol{D},\boldsymbol{M}}(\boldsymbol{m},\boldsymbol{P},\boldsymbol{\Lambda}) = \log f_{\boldsymbol{D},\boldsymbol{M}|\boldsymbol{m},\boldsymbol{P},\boldsymbol{\Lambda}}(\mathcal{D},\boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_K|\boldsymbol{m},\boldsymbol{P},\boldsymbol{\Lambda})$$
$$= \sum_{c=1}^{K}\left[\left(\sum_{i=1}^{n_c}\log\mathcal{N}\left(\boldsymbol{x}_{c,i}|\boldsymbol{\mu}_c,\boldsymbol{\Lambda}^{-1}\right)\right) + \log\mathcal{N}\left(\boldsymbol{\mu}_c|\boldsymbol{m},\boldsymbol{P}^{-1}\right)\right]$$

We can resort to the EM algorithm to perform iterative maximizations of the expected complete data log-likelihood

The EM latent variables are the class means $\boldsymbol{M}_1 \ldots \boldsymbol{M}_K$

We start from an initial estimate of the parameters — for example, we can use the data mean for $\boldsymbol{m}$, and the between and within-class covariance matrices for $\boldsymbol{P}^{-1}$ and $\boldsymbol{\Lambda}^{-1}$, respectively

In the E-step, we need to compute the posterior distributions for $\boldsymbol{M}_1 \dots \boldsymbol{M}_K$, conditioned on the current estimate of the parameters

$$f_{\boldsymbol{M}_1 \dots \boldsymbol{M}_K | \boldsymbol{D}, \boldsymbol{m}_t, \boldsymbol{P}_t, \boldsymbol{\Lambda}_t}(\boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_K | \mathcal{D}, \boldsymbol{m}_t, \boldsymbol{P}_t, \boldsymbol{\Lambda}_t)$$
$$= \prod_{c=1}^{K} f_{\boldsymbol{M}_c | \boldsymbol{D}_c, \boldsymbol{m}_t, \boldsymbol{P}_t, \boldsymbol{\Lambda}_t}(\boldsymbol{\mu}_c | \mathcal{D}_c, \boldsymbol{m}_t, \boldsymbol{P}_t, \boldsymbol{\Lambda}_t)$$

As we have seen, each term $f_{\boldsymbol{M}_c | \boldsymbol{D}_c, \boldsymbol{m}_t, \boldsymbol{P}_t, \boldsymbol{\Lambda}_t}(\boldsymbol{\mu}_c | \mathcal{D}_c, \boldsymbol{m}_t, \boldsymbol{P}_t, \boldsymbol{\Lambda}_t)$ corresponds to a Gaussian density

$$f_{\boldsymbol{M}_c | \boldsymbol{D}, \boldsymbol{m}_t, \boldsymbol{P}_t, \boldsymbol{\Lambda}_t}(\boldsymbol{\mu}_c | \mathcal{D}, \boldsymbol{m}_t, \boldsymbol{P}_t, \boldsymbol{\Lambda}_t) = \mathcal{N}(\boldsymbol{\mu}_c | \boldsymbol{m}_{t,c}, \boldsymbol{P}_{t,c}^{-1})$$

with

$$\boldsymbol{P}_{t,c} = \boldsymbol{P}_t + n_c \boldsymbol{\Lambda}_t$$
$$\boldsymbol{m}_{t,c} = \boldsymbol{P}_{t,c}^{-1} \left(\boldsymbol{\Lambda}_t \sum\nolimits_i \boldsymbol{x}_{c,i} + \boldsymbol{P}_t \boldsymbol{m}_t \right)$$

The EM auxiliary function corresponds to the expectation of the complete-data log-likelihood with respect to the posterior distribution of the latent variables

$$\mathcal{Q}(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{\Lambda}, \boldsymbol{m}_t, \boldsymbol{P}_t, \boldsymbol{\Lambda}_t)$$
$$= \mathbb{E}_{\boldsymbol{M}_1 \dots \boldsymbol{M}_K | \boldsymbol{D}=\mathcal{D}, \boldsymbol{m}_t, \boldsymbol{P}_t, \boldsymbol{\Lambda}_t} \left[\log f_{\boldsymbol{D}, \boldsymbol{M}_1 \dots \boldsymbol{M}_K | \boldsymbol{m}, \boldsymbol{P}, \boldsymbol{\Lambda}} (\mathcal{D}, \boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_K | \boldsymbol{m}, \boldsymbol{P}, \boldsymbol{\Lambda}) \right]$$

Since the likelihood and the posterior distribution factorize over classes, we can show that the auxiliary function can be expressed as

$$\mathcal{Q}(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{\Lambda}, \boldsymbol{m}_t, \boldsymbol{P}_t, \boldsymbol{\Lambda}_t)$$
$$= \sum_{c=1}^{K} \mathbb{E}_{\boldsymbol{M}_c | \boldsymbol{D}_c=\mathcal{D}_c, \boldsymbol{m}_t, \boldsymbol{P}_t, \boldsymbol{\Lambda}_t} \left[\log f_{\boldsymbol{D}_c, \boldsymbol{M}_c | \boldsymbol{m}, \boldsymbol{P}, \boldsymbol{\Lambda}} (\mathcal{D}_c, \boldsymbol{\mu}_c | \boldsymbol{m}, \boldsymbol{P}, \boldsymbol{\Lambda}) \right]$$
$$= \sum_{c=1}^{K} \mathbb{E}_{\boldsymbol{M}_c | \boldsymbol{D}_c=\mathcal{D}_c, \boldsymbol{m}_t, \boldsymbol{P}_t, \boldsymbol{\Lambda}_t} \left[\left(\sum_{i=1}^{n_c} \log \mathcal{N} \left(\boldsymbol{x}_{c,i} | \boldsymbol{\mu}_c, \boldsymbol{\Lambda}^{-1} \right) \right) + \log \mathcal{N} \left(\boldsymbol{\mu}_c | \boldsymbol{m}, \boldsymbol{P}^{-1} \right) \right]$$

Ignoring all terms that do not depend on the parameters, and denoting the expectations simply as $\mathbb{E}[\cdot]$, we have

$$
\begin{aligned}
&\mathcal{Q}(\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{\Lambda}, \boldsymbol{m}_t, \boldsymbol{P}_t, \boldsymbol{\Lambda}_t) \\
&\quad = \frac{1}{2} \sum_{c=1}^{K} \mathbb{E}\left[n_c \left|\boldsymbol{\Lambda}\right| - \operatorname{Tr}(\boldsymbol{\Lambda} \boldsymbol{S}_c) + 2 \operatorname{Tr}(\boldsymbol{\Lambda} \boldsymbol{F}_c \boldsymbol{\mu}_c^T) - \operatorname{Tr}(n_c \boldsymbol{\Lambda} \boldsymbol{\mu}_c \boldsymbol{\mu}_c^T) \right. \\
&\qquad \left. + \log \left|\boldsymbol{P}\right| - \operatorname{Tr}(\boldsymbol{P} \boldsymbol{\mu}_c \boldsymbol{\mu}_c^T) + 2 \operatorname{Tr}(\boldsymbol{m}^T \boldsymbol{P} \boldsymbol{\mu}_c) - \operatorname{Tr}(\boldsymbol{P} \boldsymbol{m} \boldsymbol{m}^T) \right] \\
&\quad = \frac{1}{2} n \log \left|\boldsymbol{\Lambda}\right| - \operatorname{Tr}(\boldsymbol{S} \boldsymbol{\Lambda}) + 2 \operatorname{Tr}\left(\boldsymbol{\Lambda} \sum_{c=1}^{K} \boldsymbol{F}_c \mathbb{E}\left[\boldsymbol{\mu}_c\right]^T \right) \\
&\qquad - \operatorname{Tr}\left(\boldsymbol{\Lambda} \sum_{c=1}^{K} n_c \mathbb{E}\left[\boldsymbol{\mu}_c \boldsymbol{\mu}_c^T\right] \right) + K \log \left|\boldsymbol{P}\right| - \operatorname{Tr}\left(\boldsymbol{P} \sum_{c=1}^{K} \mathbb{E}\left[\boldsymbol{\mu}_c \boldsymbol{\mu}_c^T\right] \right) \\
&\qquad + 2 \operatorname{Tr}\left(\boldsymbol{m}^T \boldsymbol{P} \sum_{c=1}^{K} \mathbb{E}\left[\boldsymbol{\mu}_c\right] \right) - K \operatorname{Tr}\left(\boldsymbol{P} \boldsymbol{m} \boldsymbol{m}^T \right)
\end{aligned}
$$

The terms $\boldsymbol{S}$, $\boldsymbol{S}_c$ and $\boldsymbol{F}_c$ are the statistics

$$\boldsymbol{F}_c = \sum_{i=1}^{n_c} \boldsymbol{x}_{c,i} \ , \quad \boldsymbol{S}_c = \sum_{i=1}^{n_c} \boldsymbol{x}_{c,i} \boldsymbol{x}_{c,i}^T \ , \quad \boldsymbol{S} = \sum_{c=1}^{K} \boldsymbol{S}_c$$

Setting the derivatives w.r.t $\boldsymbol{m}, \boldsymbol{P}, \boldsymbol{\Lambda}$ equal to zero we obtain

$$\boldsymbol{m}_{t+1} = \frac{1}{K} \sum_{c=1}^{K} \mathbb{E}\left[\boldsymbol{\mu}_c\right]$$

$$\boldsymbol{P}_{t+1}^{-1} = \frac{1}{K} \left(\sum_{c=1}^{K} \mathbb{E}\left[\boldsymbol{\mu}_c \boldsymbol{\mu}_c^T\right] - 2 \sum_{c=1}^{K} \mathbb{E}\left[\boldsymbol{\mu}_c\right] \boldsymbol{m}^T + K \boldsymbol{m} \boldsymbol{m}^T \right)$$

$$\boldsymbol{\Lambda}_{t+1}^{-1} = \frac{1}{n} \left(\boldsymbol{S} - 2 \sum_{c=1}^{K} \boldsymbol{F}_c \mathbb{E}\left[\boldsymbol{\mu}_c\right]^T - \sum_{c=1}^{K} n_c \mathbb{E}\left[\boldsymbol{\mu}_c \boldsymbol{\mu}_c^T\right] \right)$$

The expectations required to compute the updated model parameters can be obtained from the Gaussian density expectations. Since $\mathbb{E}\left[\boldsymbol{\mu}\boldsymbol{\mu}^T\right] = \mathbb{E}\left[\boldsymbol{\mu}\right]\mathbb{E}\left[\boldsymbol{\mu}\right]^T + Cov(\boldsymbol{\mu})$, for class mean $\boldsymbol{M}_c$, we have

$$\mathbb{E}_{\boldsymbol{M}_c|\boldsymbol{D}_c=\boldsymbol{\mathcal{D}}_c,\boldsymbol{m}_t,\boldsymbol{P}_t,\boldsymbol{\Lambda}_t}\left[\boldsymbol{\mu}_c\right] = \boldsymbol{m}_{t,c}$$
$$\mathbb{E}_{\boldsymbol{M}_c|\boldsymbol{D}_c=\boldsymbol{\mathcal{D}}_c,\boldsymbol{m}_t,\boldsymbol{P}_t,\boldsymbol{\Lambda}_t}\left[\boldsymbol{\mu}_c\boldsymbol{\mu}_c^T\right] = \boldsymbol{P}_{t,c}^{-1} + \boldsymbol{m}_{t,c}\boldsymbol{m}_{t,c}^T$$

and we can compute the parameters update as in the previous slides

We should note that, to get a robust estimate of the matrix $\boldsymbol{P}$, we require a large number of classes

Bayesian models

We can generalize the model assuming that a sample $\boldsymbol{x}_{c,i}$ of class c corresponds to the sum of three terms

$$\boldsymbol{X}_{c,i} = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{Y}_c + \boldsymbol{E}_{c,i}$$

where $\boldsymbol{U}$ is a rectangular matrix

The prior for $\boldsymbol{Y}_c$ is assumed to be

$$f_{\boldsymbol{Y}_c}(\boldsymbol{y}_c) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

and the noise terms $\boldsymbol{E}_{c,i}$ are i.i.d distributed as

$$f_{\boldsymbol{E}_{c,i}|\boldsymbol{\Lambda}}(\boldsymbol{\varepsilon}_{c,i}) = \mathcal{N}\left(\boldsymbol{\varepsilon}_{c,i}|\boldsymbol{0}, \boldsymbol{\Lambda}^{-1}\right)$$

Matrix $\boldsymbol{U}$ can be used to constrain the model means to lie in a sub-space of the original data

The model is fully specified even for rank-deficient matrix $\boldsymbol{U}$, and more robust estimates can be obtained when the number of classes is not large enough for the estimation of a full rank matrix $\boldsymbol{P}$

Bayesian models

In this case, the conditional likelihood for a sample of class c corresponds to

$$f_{\boldsymbol{X}_{c,i}|\boldsymbol{Y}_c,\boldsymbol{m},\boldsymbol{U},\boldsymbol{\Lambda}}(\boldsymbol{x}_{c,i}|\boldsymbol{y}_c,\boldsymbol{m},\boldsymbol{U},\boldsymbol{\Lambda}) = \mathcal{N}\left(\boldsymbol{x}_{c,i}|\boldsymbol{m}+\boldsymbol{U}\boldsymbol{y}_{c,i},\boldsymbol{\Lambda}^{-1}\right)$$

The derivations are similar to the previous case

The model is also known as (simplified) Probabilistic Linear Discriminant Analysis

The model we have previously seen, and the PLDA generalization, are effective when we have many classes (enough to estimate a between-class covariance matrix $\boldsymbol{P}^{-1}$, or the rectangular matrix $\boldsymbol{U}$), but few samples per class, so that we cannot reliably estimate class means

For example, this is often the case in biometric recognition systems (e.g. face, speaker or fingerprint recognition)

Bayesian models

The Bayesian model allows also addressing multi-class recognition problems

Indeed, we can model samples from the none-of-the-others class as belonging to a class $\mathcal{H}_N$ for which we have no observations

If we are using a proper prior (e.g. the one estimated with the empirical Bayes approach), we can still compute the predictive distribution for this case

The posterior distribution we use to compute the predictive density will be the prior distribution. For the Gaussian example:

$$f_{X_t|C}(\boldsymbol{x}_t|\mathcal{H}_N) = \int f_{X_t|\boldsymbol{M}}(\boldsymbol{x}_t|\boldsymbol{\mu}) f_{\boldsymbol{M}}(\boldsymbol{\mu}) d\boldsymbol{\mu} = \int \mathcal{N}\left(\boldsymbol{x}_t|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{m}, \boldsymbol{P}^{-1}\right) d\boldsymbol{\mu}$$
$$= \mathcal{N}\left(\boldsymbol{x}_t|\boldsymbol{m}, \boldsymbol{\Lambda}^{-1} + \boldsymbol{P}^{-1}\right)$$

The prior encodes the prior probability for a class mean. Since we have no observations for the class the sample belongs to, our beliefs on the corresponding class mean are encoded by the prior

Bayesian models

PLDA is an example of a more general class of models, which are known as Factor Analysis

In general, FA models express observed samples as a combination of different, Gaussian distributed factors

$$\boldsymbol{x} = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{y} + \boldsymbol{V}\boldsymbol{w} + \cdots + \boldsymbol{\varepsilon}$$

Each term $\boldsymbol{y}, \boldsymbol{w}, \ldots$ is a latent factor that models some characteristics of the data

The matrices that constrain factors to be small-dimensional are also called factor loading matrices

Usually, the residual noise term $\boldsymbol{\varepsilon}$ is assumed to have diagonal covariance matrix (PLDA is an exception, but we can represent PLDA with an alternative form which satisfies this assumption)

The factors are thus used to capture correlations between observation components

Gaussian Linear Models

Let $[\boldsymbol{x}_1 \ldots \boldsymbol{x}_n]$ be a set of (unlabeled) data. A particular instance of FA that has several practical applications consists in the model

$$\boldsymbol{x}_i = \boldsymbol{m} + \boldsymbol{W}\boldsymbol{y}_i + \boldsymbol{\varepsilon}_i$$

with

$$\boldsymbol{\varepsilon}_i \sim \boldsymbol{E}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$
$$\boldsymbol{y}_i \sim \boldsymbol{Y}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

i.e., a single factor model with isotropic noise

The conditional likelihood is given by

$$f_{\boldsymbol{X}_i|\boldsymbol{Y}_i}(\boldsymbol{x}_i|\boldsymbol{y}_i) = \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{m} + \boldsymbol{W}\boldsymbol{y}_i, \sigma^2 \boldsymbol{I})$$

Gaussian Linear Models

As for PLDA, we can compute the *posterior* distribution:

$$f_{\boldsymbol{Y}_i|\boldsymbol{X}_i}(\boldsymbol{y}_i|\boldsymbol{x}_i) = \mathcal{N}(\boldsymbol{y}_i|\boldsymbol{\mu}_{x,i}, \boldsymbol{\Lambda}_x)$$

with

$$\boldsymbol{\Lambda}_x = \frac{1}{\sigma^2}\left(\boldsymbol{W}^T\boldsymbol{W} + \sigma^2\boldsymbol{I}\right)$$
$$\boldsymbol{\mu}_{x,i} = \left(\boldsymbol{W}^T\boldsymbol{W} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{W}^T\left(\boldsymbol{x}_i - \boldsymbol{m}\right)$$

The marginal (predictive) distribution is

$$f_{\boldsymbol{X}_i}(\boldsymbol{x}_i) = \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{m}, \boldsymbol{C})$$

with $\boldsymbol{C} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}$

Gaussian Linear Models

As we did for PLDA, we can use the EM algorithm for estimating the parameters of the model

However, for this particular model closed form solutions can be given for the model parameters by direct maximization of the observed data log-likelihood

$$\sum_{i=1}^{n} -\frac{1}{2}\log|\boldsymbol{C}| - \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{m})^T \boldsymbol{C}^{-1}(\boldsymbol{x}_i - \boldsymbol{m})$$

It can be shown that a ML solution for the model is given by

$$\boldsymbol{m} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\,, \qquad \boldsymbol{W} = \boldsymbol{U}\left(\boldsymbol{L} - \sigma^2\boldsymbol{I}\right)^{\frac{1}{2}}\boldsymbol{R}\,, \qquad \sigma^2 = \frac{1}{D-M}\sum_{i=M+1}^{D} \lambda_i$$

where D and M are the dimensions of the feature and latent ($\boldsymbol{y}$) spaces, respectively

Gaussian Linear Models

$\boldsymbol{U}$ is a $D \times M$ matrix whose columns are the first M eigenvectors of the data covariance matrix

$\boldsymbol{L}$ is the diagonal matrix containing the corresponding eigenvalues λ_i for $i \leq M$

$\boldsymbol{R}$ is an arbitrary orthogonal matrix

The ML solution is not unique. In particular, we can arrange the column vectors of $\boldsymbol{U}$ so that the corresponding eigenvalues are in decreasing order

Furthermore, different choices of $\boldsymbol{R}$ correspond to the same predictive distribution[2]

Choosing $\boldsymbol{R} = \boldsymbol{I}$, the ML solution for $\boldsymbol{W}$ is then the matrix containing the first M eigenvectors of the covariance matrix, scaled by $(\lambda_i - \sigma^2)^{\frac{1}{2}}$

[2] The model is not identifiable

Matrix $\boldsymbol{W}$ thus spans the same subspace as Principal Component Analysis

Indeed, the model is known as Probabilistic Principal Component Analysis

If we assume that σ is known, and we take the limit $\sigma \to 0$, then the posterior distribution mean becomes

$$\boldsymbol{\mu}_x = \left(\boldsymbol{W}^T\boldsymbol{W}\right)^{-1}\boldsymbol{W}^T(\boldsymbol{x} - \boldsymbol{m})$$

corresponding to the PCA projection

An interesting property of the model is that we can still apply the EM algorithm even in the limit case $\sigma \to 0$

The corresponding algorithm is known as EM-PCA

EM-PCA allows computing the PCA projection for problems characterized by both a large number of samples and large sample dimensionality, for which the computation and decomposition of the data covariance matrix is unfeasible

Gaussian Linear Models

We have seen a brief introduction to Bayesian models

The main limitations of Bayesian approaches derive from the need to specify a prior (but we have seen some possible ways to select prior distributions) and to the difficulty of computing integrals for generic prior / likelihood combinations

The latter issue can be mitigated by using conjugate priors, however even simple model can quickly become intractable

Gaussian Linear Models

There exist approximate methods to handle complex posteriors, that are based on devising an approximation of the posterior distribution which makes computations tractable

If we are able to compute the posterior mode and the corresponding Hessian matrix, we can use the Laplace approximation to approximate the posterior by a Gaussian distribution

We may also specify the approximate posterior distribution as belonging to a parametric family, and look for the parameters that best approximate the intractable posterior (e.g. Variational Bayes methods)

Gaussian Linear Models

When dealing with multiple latent variables, we may assume an approximate posterior distribution that factorizes over different latent variables even in those cases where the latent variables are, a posteriori, dependent (mean-field Variational Bayes)

For further information you can refer to the course textbooks