Index:

* Variational Inference:

- Abstract: Searching an alternative to approximate probabilities densities.

- Mean field families.

- Differences respect to MCMC algorithm.

* A first application: Bayesian mixture of Gaussians
* Description of the model and the algorithm.
* Results.
* A second application: Dirichlet process for mixture model
* The workflow
* Results.

1. Variational Inference
   1. Abstract: searching an alternative to approximate probabilities densities:

One of the core problems of modern statistics is to approximate difficult-to-compute probability densities. This problem is especially important in Bayesian statistics, which frames all inference about unknown quantities as a calculation involving the posterior density. We are going to discuss about Variational inference (VI), a deterministic method that approximates probability densities through optimization which is faster than the classical approaches such as Markov Chain Montecarlo sampling (MCMC). The idea behind VI is to first posit a family of densities and then to find the member of that family which is close to the target. Closeness is measured by Kullback-Leibler divergence.

The first step is to specify a family *Q* of densities over the latent variables **z**. Each *q in Q* is a possible approximation of the real density. Our goal is to find the one candidate, which minimizes the KL divergence. Inference now amounts to solving the following optimization problem:

Immagine che contiene Carattere, testo, calligrafia, bianco

Descrizione generata automaticamente

But this object is not directly computable because looking the full expression:

is not possible to compute the last term. So, we optimize an alternative object which is equivalent to KL up to an added constant:



This expression is called evidence lower bound (ELBO) and since it corresponds to the negative KL, we are going to maximise it.

We rewrite the ELBO as a sum of the expected log likelihood of the data **x** and the KL divergence between the prior *p*(**z**) and *q*(**z**):

Immagine che contiene testo, Carattere, tipografia

Descrizione generata automaticamente

The first term is an expected likelihood; it encourages densities that place their mass on configurations of the latent variables that explain the observed data. The second term is the negative divergence between the variational density and the prior; it encourages densities close to the prior.

Another property of the ELBO is that it lower-bounds the (log) evidence, log *p*(**x**) *≥* ELBO(*q*) for any *q*(**z**). This explains the name. To see this look the following expression:



The bound then follows from the fact that KL (*·*) *≥* 0.

* 1. The mean-field variational family

We now describe a variational family *Q*, to complete the specification of the optimization problem. The complexity of the family determines the complexity of the computations.

Here we give a glimpse of the *mean-field variational family*: the latent variables are mutually independent and each governed by a distinct factor in the variational density. A generic member of the mean-field variational family is

Immagine che contiene Carattere, bianco, testo, tipografia

Descrizione generata automaticamente

Each latent variable  *z\_j* is governed by its own variational factor, the density *q\_j*(*z\_j*). We emphasize that the variational family is not a model of the observed data, indeed, the data do not appear in the previous equation. Instead, it is the ELBO, and the corresponding KL minimization problem, that connects the fitted variational density to the data and model.

* 1. Differences respect to Montecarlo Markov chains methods:

The main advantage of the VI respect to MCMC is the capability to assess the solution faster in terms of the time. This creates the possibility explore an important amount of possible differences solutions. In other terms this methos is perfectly suitable for large datasets and scenarios where to explore many models.

But the price that we pay for this is that the solution is not exact.

In the other hand MCMC algorithms are very slow in terms of time and very heavy in terms of computations. But asymptotically give a precise solution. This methos is perfect for small datasets and more precise samples are needed.

So, there is not a right or wrong method. The choice depends on the goal we want to achieve. But one thing is clear: we are going through a time/exact solution trade-off.

1. A first application: Bayesian mixture of Gaussians
   1. Description of the model and the algortitm:

We are going to considerer a Bayesian mixture of multivariate Gaussians which has K components, corresponding to K Gaussian distributions with means vectors *{µ\_*1,.. ,*µ\_K}.*

We assume that the data have dimension d.

The mean parameters are independent from each other and we assume that they are distributed as Gaussian *N* (0, (*σ^*2)\*I), where the prior variance *σ^*2 is a hyperparameter. To generate an observation *xi* from the model, we first choose a cluster assignment *ci*. It indicates which latent cluster *xi* belongs and is distributed as a categorical distribution over {1, …, *K}*.

Note that ci is a K-dimension vector with all zeros except in the position of xi’s cluster.

Finally the datum xi has gaussian distribution of *N* ((*ci^T)\*****µ***,I).

Here the complete hierarchical model:

*Immagine che contiene testo, Carattere, bianco, linea

Descrizione generata automaticamente*

In this case the mean-field variational family components have the following form:

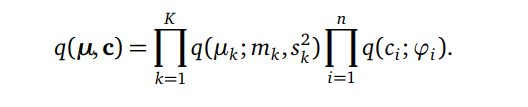


Immagine che contiene testo, Carattere, schermata, linea

Descrizione generata automaticamentewhere:

So hour focus is to compute the following parameters:

Immagine che contiene testo, Carattere, schermata, bianco

Descrizione generata automaticamente

**Immagine che contiene testo, Carattere, schermata, bianco

Descrizione generata automaticamente**The ELBO expression that we are going to maximise is in function of the variational parameters **m, s^2** and **phi,**

Note that each term can be computed in closed form. Let’s go deeper about it:

* Immagine che contiene Carattere, testo, bianco, calligrafia

  Descrizione generata automaticamenteFirst term is equal to:
* The second is -N\*log(K) that can be discarded since is constant.
* Immagine che contiene testo, Carattere, bianco, Elementi grafici

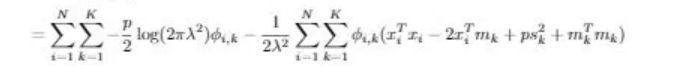
  Descrizione generata automaticamenteThe third term:
* First notice that:
* So:

Immagine che contiene testo, Carattere, bianco, design

Descrizione generata automaticamente

* The fourth:

Immagine che contiene testo, Carattere, bianco, logo

Descrizione generata automaticamente

* The fifth:

Togliere alpha dall’espressione e sistemare notazioni opportune

The variational parameter are updated in turn. Here the expressions:

* Cluster assignment:
* Mixture component means :



And if we pass to the logarithm there is the following computations

Immagine che contiene testo, Carattere, bianco, algebra

Descrizione generata automaticamente

This calculation reveals that the coordinate-optimal variational density of *µ\_k* is an exponential family with sufficient statistics *{(µk)*,(*µ*\_k)^2*}* and natural parameters

i.e a Gaussian.

Immagine che contiene Carattere, testo, linea, diagramma

Descrizione generata automaticamenteExpressed in terms of the variational mean and variance, the updates for q(u\_k) are

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamenteHere there is the pseudocode of the algorithm that we implemented in Python:

* 1. Results:

After gathering the results we have implemented the Gibbs Sampler algorithm always in Python to discover the points of strength and weakness of VI:

The table below describe everything:

Immagine che contiene testo, schermata, numero, Carattere

Descrizione generata automaticamenteConsiderer dimension d = 2 and number of cluster k = 5. While N is the sample size.

Qui ci sta scrivere le metriche di errore piu nel dettaglio.

The table confirms that VI is much faster than MCMC but gives a less precise solution. While the clustering quality is practically the same.