Index:

* Variational Inference:

- Abstract: Searching an alternative to approximate probabilities densities;

- Mean field families;

- Differences respect to MCMC algorithm;

* A first application: Gaussian mixture model
* 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. 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:

but 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 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 *qj*(*zj*). We emphasize that the variational family is not a model of the observed data—indeed, the data **x** does 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 asses to 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 in large datasets and scenarios where to explore many models.

The big disadvantage is that the solution is not exact.

In the other hand MCMC algorithms are very slow in terms of time and very heavy in computation. 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 from the goal we want to achieve. But one thing is clear: we are going through a time/exact solution trade-off.