Fisher metric. A natural gradient ascent: Expectation maximization

Contents

Introduction

Suppose that we are trying to fit a probabilistic model $p_\theta(x)$ to data $x_1, ..., x_n$ by a gradient ascent (our global problem is finding $\text{argmax}_{\theta\in\Theta}(p_\theta(x))$.

As we showed in the two previous talks, the gradient ascent is of the form

$$\theta \leftarrow \theta + \eta M^{-1}\frac{\partial L}{\partial \theta}, \tag{1}$$

where L is the loss function we are trying to minimize (a relevant choice for $L(\theta)$ could be the length of the data compressed using p_θ, for example), and M is a metric on Θ. We were hoping to find a "reasonable" metric M in the sense that it should not depend on the parametrization.

We also recall that the gradient ascent can be rewritten (up to $O(\eta^2)$) as the following penalized maximization problem:

$$\theta^{n+1} = \text{argmax}_\theta\{f(\theta) - \frac{1}{2\eta}\|\theta - \theta^n\|_M^2\}, \tag{2}$$

where $\|a\|_M^2 = a^T M a$. M can therefore be seen as a penalty for moving away from the current point. In the case of the probabilistic model, we want this penalty to depend only on p_θ and p_{θ^n}. One way to do this is to use the Kullback–Leibler divergence.

1 Kullback–Leibler divergence, Fisher metric

For the remainder of the text, we will suppose that the sample space X is discrete. If X is continuous, the sums simply have to be replaced by integrals.

1.1 Kullback–Leibler divergence

Consider an emitter sending data following the probability distribution p, and suppose the receiver knows p, and tries to encode the data.

To minimize the length of the encoded message, the receiver should encode x with $-\log_2 p(x)$ bits, yielding the codelength $H(p) = -\sum p(x) \log_2 p(x)$, which is by definition the Shannon entropy of p.

Now, if the receiver does not know p, and uses a probability distribution q instead to encode the data, the codelength will be $-\sum p(x) \log_2 q(x)$.

By definition, the Kullback–Leibler divergence (already introduced in the first talk) is the difference between these two codelengths:

$$\mathrm{KL}(p\|q) := \sum p(x) \log_2 \frac{p(x)}{q(x)}, \tag{3}$$

which is always positive. In other words, it is best to use the probability distribution p to compress data generated with p.

The Kullback–Leibler divergence is not symmetric, and is hard to manipulate in practice. The reasonable thing to do is to use the second order approximation of the Kullback–Leibler divergence: the Fisher metric.

1.2 Fisher metric

By definition, the Fisher metric is the second order term in $\delta\theta$ of $\mathrm{KL}(p_{\theta+\delta\theta}\|p_\theta)$: it is possible to show that

$$\mathrm{KL}(p_{\theta+\delta\theta}\|p_\theta) = \frac{1}{2}\delta\theta^T I(\theta)\delta\theta + o(\delta\theta^2), \tag{4}$$

where

$$I(\theta) := E_{x\sim p_\theta}\left[-\frac{\partial^2 \ln p_\theta(x)}{\partial\theta^2}\right] = E_{x\sim p_\theta}\left[\frac{\partial \ln p_\theta(x)}{\partial\theta}^T \frac{\partial \ln p_\theta(x)}{\partial\theta}\right] \tag{5}$$

is the so-called Fisher metric.

We can now write the corresponding gradient ascent, called *natural gradient ascent*, whose advantages will be discussed in the remainder of the section:

$$\theta \leftarrow \theta + \eta I(\theta)^{-1} \frac{\partial L}{\partial\theta}, \tag{6}$$

which is *invariant with respect to the parametrization*, since it has been defined using only the p_θ and not the θ.[1]

The Fisher information also has an interpretation linking it to the "precision" of estimators.

[1] It is known that the Fisher metric is the only invariant metric for probability distributions which has certain "reasonable" properties.

1.2.1 Cramer–Rao bound

Intuitively, if $\{p_\theta, \theta \in \Theta\}$ is a family of probability distributions, an estimator on Θ takes observations $x_1, ..., x_n$ as an argument, and returns a value of the parameter θ which is believed to be the value used to sample the x_i. Saying an estimator is unbiased means that its expected value is the parameter θ used to sample the data.[2] This yields the following definition.

Definition 1. *Let X be a set, $\{p_\theta, \theta \in \Theta\}$ a family of probability distributions on X.*

An estimator *on Θ is a function from X^* to Θ.*

An estimator is said to be unbiased *if $\forall \theta \in \Theta, E_{x \sim p_\theta} \psi(x_1, ..., x_n) = \theta$*

We can now write the Cramer–Rao bound, which is a lower bound on the variance of unbiased etimators.

Theorem 2 (Cramer–Rao bound)**.** *If $x_i \sim p_\theta$, and if ψ is an unbiased estimator on Θ, then*

$$\mathrm{Var}(\psi(x_1, ..., x_n)) \geqslant \frac{1}{n} I^{-1}(\theta). \tag{7}$$

The proof can be found in [2].

In other words, the Fisher matrix at θ defines a "box"[3] around θ such that it is not possible to know if the data has been sampled from θ or from some θ' in the box. For example, if the Fisher information at θ is large, a small variation of θ will yield a large variation of p_θ, and consequently, the box will be smaller.

An important result concerning natural gradient is that an estimator trained with natural gradient asymptotically reaches the Cramer-Rao bound (Theorem 2 in [1]).

1.2.2 Relationship with diversity

Another interesting point with the natural gradient ascent is that the Kullback–Leibler divergence penalizes the loss of diversity. For example, if we start with a uniform distribution, we find:

$$\mathrm{KL}(p_\theta \| \mathrm{Unif}) = \mathrm{cst} - H(p_\theta), \tag{8}$$

and entropy is a reasonable way of measuring diversity. This remark is important in machine learning, because keeping a high diversity should prevent overfitting.

We are now going to present the expectation maximization algorithm, an algorithm used for prediction with missing data that can be described as a natural gradient ascent.

2 Expectation maximization

Suppose we have data x, with $x = (x', x'')$, but x'' is missing, and consider a family $\{p_\theta, \theta \in \Theta\}$ of probability distributions, from which we would like to pick the "best" θ to explain the data.

[2] Notice the important assumption that the data are sampled from some p_θ.

[3] The box is the ellipsoid defined by the equation $(x - \theta)^T \frac{1}{n} I(\theta)(x - \theta) \leqslant 1$

The expectation maximization algorithm maintains an estimate θ_n updated as follows: firstly, try to synthesize the missing data x'' according to $p_{\theta_n}(x''|x')$, thus obtaining an estimated dataset $\hat{x}$, and then, estimate θ_{n+1} based on $\hat{x}$.

There are several possibilities for the estimation of the dataset.

1. Pick one value of x'' at random from $p_{\theta_n}(x''|x')$.
2. Pick the most probable value of x'' from $p_{\theta_n}(x''|x')$.
3. Pick a few values of x'' weighted by $p_{\theta_n}(x''|x')$.
4. Pick all possible values x'' weighted by $p_{\theta_n}(x''|x')$.

We obtain the following algorithms:

Definition 3 (Expectation maximization algorithm)**.** *We define the four following update rules for θ_n.*

1. *Synthesize x'' using θ_n (i.e. $x'' \sim p_{\theta_n}(.|x')$).*
   Maximize $\theta_{n+1} := \operatorname{argmax}_\theta \ln p_\theta(\hat{x})$
2. *Pick the most probable x'' knowing x', using θ_n (i.e. $x'' = \operatorname{argmax}_z p_{\theta_n}(x', z)$).*
   Maximize $\theta_{n+1} := \operatorname{argmax}_\theta \ln p_\theta(\hat{x})$
3. *[Monte-Carlo expectation maximization]*
   Take k samples $x''_1, ..., x''_k$, with $x''_j \sim p_{\theta_n}$ for $j \in [1, k]$, and give each of these samples a weight $\frac{1}{k}$.

$$\theta_{n+1} = \operatorname{argmax}_\theta \frac{1}{k} \sum_{j=1}^{k} \ln p_\theta(x', x''_j). \tag{9}$$

4. *[Classical expectation maximization]*

$$\theta_{n+1} = \operatorname{argmax}_\theta \sum_{x''} p_{\theta_n}(x''|x') \ln p_\theta(x', x''). \tag{10}$$

   This is the limit of (9) for $k \to \infty$.

   Notice that only variants 2 and 4 are deterministic.

Expectation maximization can be used for example for clustering, where the x'_i are points, and x''_i is the cluster to which x'_i belongs (variant 2 is the k-means algorithm). It is also used for hidden Markov models, for which (10) can be computed exactly.

2.1 Link with gradient ascent

The expectation maximization algorithm can be linked with gradient descent as follows: let us define the function $L(\theta) := \ln p_\theta(x') = \ln \sum_{x''} p_\theta(x', x'')$, which is what we are trying to maximize (it is the probability of seeing the data we actually observed).

Proposition 4. *The EM algorithm* (10) *is equivalent to setting:*

$$\theta_{n+1} = \text{argmax}_\theta \{L(\theta) - \text{KL}\left(p_{\theta^n}(x''|x') \| p_\theta(x''|x')\right)\} \tag{11}$$

Proof. We start from (11). We have:

$$\begin{aligned}
\theta_{n+1} &= \text{argmax}_\theta \left\{L(\theta) - \text{KL}(p_{\theta^n}(x''|x') \| p_\theta(x''|x'))\right\} \\
&= \text{argmax}_\theta \left\{ \ln p_\theta(x') - \sum_{x''} p_{\theta^n}(x''|x') \ln \frac{p_{\theta^n}(x''|x')}{p_\theta(x''|x')} \right\} \\
&= \text{argmax}_\theta \left\{ \ln p_\theta(x') + \sum_{x''} p_{\theta^n}(x''|x') \ln p_\theta(x''|x') \right\}
\end{aligned}$$

(since the removed term $p_{\theta^n}(x''|x') \ln p_{\theta^n}(x''|x')$ does not depend on θ)

$$\begin{aligned}
&= \text{argmax}_\theta \left\{ \sum_{x''} p_{\theta^n}(x''|x') \ln p_\theta(x') + \sum_{x''} p_{\theta^n}(x''|x') \ln p_\theta(x''|x') \right\} \\
&= \text{argmax}_\theta \left\{ \sum_{x''} p_{\theta^n}(x''|x') \left(\ln p_\theta(x''|x') + \ln p_\theta(x')\right) \right\} \\
&= \text{argmax}_\theta \left\{ \sum_{x''} p_{\theta^n}(x''|x') \ln p_\theta(x', x'') \right\},
\end{aligned}$$

which is exactly (10). □

As an immediate corollary, the EM algorithm improves L at each step (at worst, take $\theta_{n+1} = \theta_n$).

The link with natural gradient is now clear: (11) is essentially (2), with step size $\eta = \frac{1}{2}$, and with the KL divergence *on* x'' instead of the Fisher metric.

References

[1] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[2] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Elements of Information Theory. Wiley, 2006.