Two approaches to kernel construction

Seminar report for the course T-122.102

Leo Lahti April 16, 2004

Introduction

The notion of similarity is a fundamental concept in data analysis. Similarity can be defined in a number of ways and is often mathematically encoded by a distance function as a *metric*. Close-by objects in a given metric are similar whereas more distant objects are dissimilar. Useful metric depends on the problem.

Kernels are functions that can be used to encode distance information between objects. It is, however, not a straightforward task to encode our notion of similarity as kernel. For intelligent representations of distance information via kernels, we need principled ways to choose a suitable kernel for a given purpose. It can be picked up from a selection of standard kernels or, even better, specifically designed to suit well the given purpose.

This seminar presentation is based on two articles [3, 4] that work kernel construction methods based on uncompletely known data sets. In the first approach, marginalized kernels are derived from latent variable models. In the second approach, a kernel is designed from unlabeled data using mutual information framework.

Both approaches have a connection with the widely used Fisher kernel [1].

Marginalized kernels

The goal of marginalized kernel approach is to design a kernel from a latent variable model where we have both visible and hidden variables that are both relevant for the modeling task. An example of such are biological sequences for which we have visible variables, namely the sequence items, and context information for the items that is typically hidden. Here the context information indicates for example whether the given sequence item is in a coding or a non-coding region of the sequence.

Marginalized kernels are derived in two steps. First, a *joint kernel* $K_z(z, z')$ for the whole model with visible (x, x') and hidden (h, h') variables is constructed. Here $z = (x, h)$ and $z' = (x', h')$ denote the joint variables of the model. In the second step, we take expectation of the joint kernel over hidden variables h, h', resulting in a *marginalized kernel*

$$K(x, x') = \sum_{h,h'} p(h|x)p(h'|x)K_z(z, z').$$

that is readily usable for visible variables only. Conditional posterior distributions $p(h|x)$ and $p(h'|x')$ for the hidden variables are unknown in general and have to be estimated using suitable models, e.g. hidden Markov models. It needs to be emphasized that a selection of a suitable joint kernel depends on the problem.

Practical example

In text processing literature, sequence comparisons have been succesfully performed using the *count kernel*:

$$K(x, x') = \sum_k c_k(x) c_k(x'),$$

$$c_k(x) = \frac{1}{m} \sum_i I(x_i = k),$$

where $c_k(x)$ is the percentage of occurences of symbol k in sequence x. For biological sequences this not enough, however- we need also context information. Sequence items have different meaning depending on their coding status.

DNA structure can be represented by four symbols (A,C,G,T nucleotides). DNA is divided in coding and non-coding regions and this information is typically hidden for genomic sequences. Let us now study a genomic sequence x and an indicator variable h with 1 and 2 denoting coding and non-coding region, respectively, for the corresponding nucleotide:

$$\mathbf{h} = 122122122$$
$$\mathbf{x} = \text{ACGGTTCAA}$$

We represent the joint kernel for visible nucleotide and hidden context information as

$$K_z(z, z') = \sum_k \sum_l c_{kl}(z) c_{kl}(z'),$$
$$c_{kl}(z) = \frac{1}{m} \sum_i I(x_i = k, h_i = l).$$

This has apparently similar form than the usual count kernel described earlier. So, in the DNA case we apply count kernel for the sequence model with both the sequence and the context information. Unfortunately, the context information is hidden and have to be marginalized out. Marginalizing the joint kernel over hidden variables we get

$$K(x, x') = \sum_{\mathbf{h}, \mathbf{h}'} p(\mathbf{h}|x) p(\mathbf{h}'|x) K_z(z, z') = \sum_{k,l} \gamma_{kl}(x) \gamma_{kl}(x'),$$

where

$$\gamma_{kl}(x) = \frac{1}{m} \sum_i \sum_{h_i} p(h_i|x) I(x_i = k, h_i = l).$$

Unknown conditional posteriors $p(h_i|x)$ can be estimated in a straightforward manner using hidden Markov models (see [4]).

Evaluation of marginalized kernels

The joint kernel $K(z, z')$ should be designed for the given purpose. The advantage of the marginalization approach is that the joint kernel and the probabilistic model $p(x|\theta)$ can be completely separated. This allows utilizing higher order information with a first order HMM, for example, contextual dependencies for consecutive sequence items.

Marginalization approach is especially useful when context information is crucial as is often case for biological sequences. It was also shown that the marginalized count kernel outoutperformed the Fisher kernel in bacterial classification. The authors plan to derive a variety of kernels for practical purposes in their future works using the marginalization approach.

Mutual information for learning covariance kernels

We now proceed to the second approach for constructing a kernel. As mentioned, kernel should somehow encode our notion of similarity. Clusters in probability distribution $P(x)$ contain similar samples, and samples in different clusters are dissimilar. Mutual information (MI) between the samples is one way to represent such similarities, and it can be computed for unlabeled data.

This leads the author in [3] to suggest deriving a kernel using the concept of mutual information as starting point. The aim is to learn covariance kernel from unlabeled data set. Covariance kernels are a subclass of kernels with mathematically nice features, like positive definiteness.

Construction of a MI kernel proceeds in the following steps. Choose a model family $\{P(\mathbf{x}|\theta)\}$ and a prior distribution $P(\theta)$ for the parameters θ. Then fit generative models to unlabeled data D_u using the estimation

$$P(\theta|D_u) \propto P(D_u|\theta)P(\theta).$$

Relying too much on unlabeled information in the kernel construction process may result in the lack of robustness of the final results. We can adjust the effect of unlabeled data in the kernel learning process by adjusting λ parameter in *model-trust scaling.*

$$P_{med}(\theta|\lambda) \propto P(D_u|\theta)^{\lambda/n}P(\theta), \lambda \in [0, n].$$

Here the λ defines how much the original data is given weight w.r.t. the prior distribution. Finally we can build a covariance kernel using this posterior information. Let us define the joint distribution between two data points as

$$Q(x_1, x_2) := \int P_{med}(\theta)P(x_1|\theta)P(x_2|\theta)d\theta,$$

where $P_{med}(\theta)$ is a *mediator distribution.* From the joint distribution we derive *MI-score*:

$$I(x_1, x_2) = log\frac{Q(x_1, x_2)}{Q(x_1)Q(x_2)},$$

where

$$Q(x) = \int Q(x, x')dx'.$$

In a very concrete sense, it measures the similarity between x_1 and x_2 with respect to the generative process represented by the mediator distribution. It is the amount of information they share via the mediator variable θ. However, this is not *positive definite* as is required for a proper covariance kernel. To get a positive definite kernel from MI-score, we use exponential embedding:

$$\begin{aligned} K(x_1, x_2) &= exp(-\frac{1}{2}[I(x_1, x_1) + I(x_2, x_2)] + I(x_1, x_2)) \\ &= \frac{Q(x_1, x_2)}{\sqrt{Q(x_1, x_1)Q(x_2, x_2)}}. \end{aligned}$$

This is called the *MI kernel.* The exponential embedding trick becomes familiar by noting that it converts the usual inner product to radial basis functions (see [3]).

The derivation of a covariance kernel using mutual information proceeds in a rather intuitive manner. However, the practical examples that are represented in the paper are not very promising and some extensions to the basic framework seem necessary.

Connections to Fisher kernel

Both kernel construction approaches have a connection with the Fisher kernel [1] that is widely used in kernel-based data analysis. Fisher kernel K_f is defined by

$$K_f(x,x') = s(x,\theta')^T Z^{-1}(\theta')s(x,\theta'),$$

with the Fisher score

$$s(x,\theta') = \nabla_\theta logp(x|\theta')$$

and the Fisher information matrix

$$Z(\theta') = E[s(x,\theta')s(x,\theta')^T|\theta'].$$

Fisher kernel as a special case of marginalized kernels

Let us adopt a latent variable model $p(x|\theta) = \sum_h p(x,h|\theta)$. The Fisher score now takes the form

$$\begin{aligned} s(x,\theta') &= \nabla_\theta logp(x|\theta') \\ &= \nabla_\theta log \sum_h p(x,h|\theta) \\ &= \sum_h p(h|x,\theta')\nabla_\theta logp(x,h|\theta'). \end{aligned}$$

Then the corresponding Fisher kernel is

$$\begin{aligned} K_f(x,x') &= s(x,\theta')^T Z(\theta')^{-1}(\theta')s(x,\theta') \\ &= \nabla_\theta logp(x|\theta')^T Z(\theta')^{-1}\nabla_\theta logp(x|\theta') \\ &= \sum_{h,h'} p(h|x,\theta')p(h'|x',\theta')K_z(z,z') \end{aligned}$$

We notice that for the Fisher kernel, the joint kernel is described as

$$K_z(z,z') = \nabla_\theta logp(x,h|\theta')^T Z(\theta')^{-1}\nabla_\theta logp(x',h'|\theta').$$

So the Fisher kernel is a special case of marginalized kernels, with the above joint kernel. This adds a new aspect to the Fisher kernel theory.

Fisher kernel as a special case of MI-derived covariance kernels

For mutual information approach, we are also able to derive Fisher kernel as a special case, assuming flat prior distribution $P(\theta) = 1$ and the *Laplace approximation* for the MI-derived ovariance kernel (see e.g. [2]). With different approximations, the MI based kernel has connections to other kinds of kernels as well (see [3]).

Conclusions

It is not easy to encode our notion of similarity as a kernel. This is problem especially when data is only partially known. At such situation, kernel learning approaches may be useful. Two approaches for kernel construction for hidden varible models and from unlabeled data sets were presented.

It is also noted that both approaches have a connection with a Fisher kernel. The marginalized kernel approach offers a new aspect on the Fisher kernel theory, deriving the Fisher kernel as a special case of marginalized kernels. For mutual information kernel approach, the Fisher kernel appears with suitable approximations and prior assumptions.

References

[1] T. Jaakkola, M. Diekhans and D. Haussler, "Using the Fisher kernel method to detect remote protein homologies," in **Proceedings of the 7th Annual International Conference on Intellegent Systems for Molecular Biology (ISMB'99)**, Menlo Park, CA: AAAI Press, pp. 1149–1158, 1999.

[2] M. Seeger, "Covariance kernels from Bayesian generative models," Techn. report, **Institute for ANC**, Edinburgh, UK, 2000.

[3] M. Seeger, "Covariance kernels from Bayesian generative models," in T. Dietterich, S. Becker and Z. Ghahramani (eds.), **Advances in Neural Information Processing Systems 14**, Cambridge, MA: MIT Press, pp. 905–912, 2002.

[4] K. Tsuda, M. Kawanabe, G. Rätsch, S. Sonnenburg and K.-R. Müller, "A new discriminative kernel from probabilistic models," **Neural Computation**, vol. 14, pp. 2397–2414, 2002.