Bayesian classification with discrete templates

prepared by: Coryn A.L. Bailer-Jones[1], David W. Hogg[1,2]
affiliation : [1]MPIA, [2]NYU

reference: GAIA-C8-TN-MPIA-CBJ-061
issue: 1
revision: 1
date: 2011-08-29
status: Issued

Abstract

We outline the principles of Bayesian classification using discrete templates. We show that this form of nonparametric modelling is semi-supervised learning and embodies multi-level hierarchical modelling to infer all model parameters from the data.

Document History

Issue	Revision	Date	Author	Comment
1	1	2011-08-29	CBJ	Added softmax implementation
1	0	2011-08-21	CBJ	Changes following discussions with DWH
D	0	2011-08-17	CBJ	Added amplitudes; discussions with DWH
D	0	2011-07-25	CBJ	First draft

Contents

TABLE 1: Notation

k	class label, $k = 1 \dots K$
$T_j^{(k)}$	data (vector) on template j of class k, $j = 1 \dots J_k$
$c_j^{(k)}$	amplitude of template j of class k
x_n	measured data (vector), for object n. $\{x_n\}$ is the set of all x_n
V_n	covariance matrix of x_n. $\{V_n\}$ is the set of all V_n
$\beta_j^{(k)}$	$= P(T_j^{(k)} \vert k)$, template relevance parameter
$\{\beta_j^{(k)}\}_j$	set of all template relevances for class k
$\{\beta_j^{(k)}\}$	set of all template relevances for all classes
α_k	$= P(k \vert \{\beta_j^{(k)}\}_j)$, class relevance parameter
$\{\alpha_k\}$	set of all class relevance parameters

1 The basic idea

Context: We have a set of N measured objects each with data x_n (in general a vector, such as a spectrum), $n = 1 \dots N$. We wish to classify these in a system of K classes, $k = 1 \dots K$. Each class is represented by a set of J_k templates, $j = 1 \dots J_k$, for which we have $x = T_j^{(k)}$ for each. These templates could be empirical (labelled data) or synthetic, but they are taken "as is".

Objective: The main goal is to estimate the class posterior probabilities $P(k|x_n, \dots)$, where $\dots$ indicates some other properties of the data (e.g. measurement uncertainties) and parameters inferred from the data (plus assumptions).[1]

As we only have a discrete set of templates, we must assume that each measured object is actually a noisy measurement of one of the templates. This is likely to be a poor assumption, but it is the limitation imposed by the decision to use discrete templates. Given sufficient reason we could later add more templates or more classes. Let us adopt a model for this noise (*likelihood model*) , the usual choice being the Gaussian[2]

$$P(x_n|T_j^{(k)}, k, V_n) = \frac{1}{(2\pi)^{I/2}|V_n|^{1/2}} \exp\left[-\frac{1}{2}\left(x_n - T_j^{(k)}\right)^{\mathsf{T}} V_n^{-1} \left(x_n - T_j^{(k)}\right)\right] \tag{1}$$

where V_n is the covariance of the measurement x_n, $^{\mathsf{T}}$ indicates transpose, and I is the dimensionality of the data vectors. (In the case of uncorrelated noise, V_n reduces to a diagonal matrix with elements equal to the "measurement uncertainties", and the equation can be written as a product of 1D Gaussians.)[3]

[1] In terms of publishing classification results from a survey, we may also want to publish the likelihoods $P(k|\{x_n\}, \dots)$, as this makes it easier for users to apply their own priors.

[2] This is inappropriate if the data are strictly positive, but for flux measurements obtained after background subtraction (which can therefore be negative) it may be an adequate approximation.

[3] We could also use equation 1 to introduce noise in the templates, or more generally, some smoothing/tolerance

The probability that x_n is a noisy measurement of *any* of the templates in class k is

$$P(x_n|k, V_n) \;=\; \sum_{j=1}^{j=J_k} P(x_n|T_j^{(k)}, k, V_n) P(T_j^{(k)}|k) \tag{2}$$

which follows from the rules of probability (and noting that the second term in the sum has no dependence on V_n). The terms $P(T_j^{(k)}|k)$ describe the distribution of the class templates. If we had a continuous model for the data, then $P(T_j^{(k)}|k)$ would be replaced by $P(T^{(k)}(w)|k)dw$ and the summation replaced by an integral over the parameters, w, of the continuous model. In the present discrete case, $P(T_j^{(k)}|k)$ has no distribution as such because $T_j^{(k)}$ has no parameters. Each $P(T_j^{(k)}|k)$ is just a scalar number: the probability of a template. We consider these as parameters of the model and write them as $\beta_j^{(k)}$. We then write equation 2 as

$$P(x_n|k, \{\beta_j^{(k)}\}_j, V_n) \;=\; \sum_{j=1}^{j=J_k} P(x_n|T_j^{(k)}, k, V_n)\, \beta_j^{(k)} \tag{3}$$

where $\forall j, k$: $0 \leq \beta_j^{(k)} \leq 1$ and $\forall k : \sum_j \beta_j^{(k)} = 1$. Note that we have now introduced an explicit dependence on these parameters into the likelihood on the LHS.

In principle we could specify $\{\beta_j^{(k)}\}$ a priori. We may set them equal, for example, reflecting a prior belief that they are all equally good representatives for their class. But more generally we treat them as free parameters and infer them from the data. That is, we maximize (something involving) $P(x_n|k, \ldots)$ with respect to them. This gives us the model, $\{\beta_j^{(k)}\}$, which makes the data most probable. (How to do this is outlined in section 2.) Note that there is one set of these parameters for the whole data set. As they are template probabilities, constrained to the range 0–1, we interpret them as *template relevances*. If the templates are broadly distributed in the data space w.r.t the x_n and if the variances in V_n are comparatively small, then many of these probabilities will be negligible. Those templates which best explain the most data achieve higher values of $\beta_j^{(k)}$, subject to the constraints. Thus increased relevance of one template must come at the cost of reduced relevance of one or more of the others.

Once $\{\beta_j^{(k)}\}$ have been found, the degree to which each template is relevant for modelling an individual measurement, x_n, is determined by the likelihood term in equation 3. We can view that equation as a β-weighted mixture model of J_k Gaussians.

In the continuous model case we would train the model to best fit the training data, i.e. to make the best predictions of these data. In the present case, in contrast, we use both the templates *and* the unlabelled data to optimize these parameters. In machine learning speak, we are also using the test data to optimize the model parameters. Hence this discrete classification method is a case of semi-supervised learning rather than pure supervised learning.

parameter for how well we require a template to reproduce a measured object. This too could be inferred from the data.

We can view this discrete model as a "non-parametric" model, consisting of a set of delta functions in the data space (although the term "non-parametric" is really an oxymoron, because this model actually has one parameter per template, which is many more than a smooth continuous model).

Having found $P(x_n|k)$ for all n and k, the final step is to use Bayes' theorem to get the posterior probability for each class

$$\begin{aligned} P(k|x_n, \{\beta_j^{(k)}\}_j, V_n) &= \frac{P(x_n|k, \{\beta_j^{(k)}\}_j, V_n) P(k|\{\beta_j^{(k)}\}_j, V_n)}{P(x_n|\{\beta_j^{(k)}\}_j, V_n)} \\ &= \frac{P(x_n|k, \{\beta_j^{(k)}\}_j, V_n)\, \alpha_k}{\sum_k P(x_n|k, \{\beta_j^{(k)}\}_j, V_n)\, \alpha_k} \end{aligned} \tag{4}$$

where $\alpha_k = P(k|\{\beta_j^{(k)}\}_j)$, in which the dependence on V_n has been dropped because it provides no additional information given $\{\beta_j^{(k)}\}_j$ (conditional independence). These α terms are also parameters of the model, the *class relevances*. They are subject to the constraints $\forall k$: $0 \leq \alpha_k \leq 1$ and $\sum_k \alpha_k = 1$. These too can be inferred from the data, as we will now see.

2 Inferring the relevance parameters

For a given set of data, $\{x_n\}$, we determine a single set of $\beta_j^{(k)}$ parameters. To infer these we maximize $P(\{x_n\}|\{\beta_j^{(k)}\}, \{V_n\})$ w.r.t these, where

$$\begin{aligned} P(\{x_n\}|\{\beta_j^{(k)}\}, \{V_n\}) &= \prod_n P(x_n|\{\beta_j^{(k)}\}, V_n) \\ &= \prod_n \sum_k P(x_n|k, \{\beta_j^{(k)}\}, V_n) P(k|\{\beta_j^{(k)}\}, V_n) \\ P(\{x_n\}|\{\beta_j^{(k)}\}, \{\alpha_k\}, \{V_n\}) &= \prod_n \sum_k P(x_n|k, \{\beta_j^{(k)}\}_j, V_n)\, \alpha_k \end{aligned} \tag{5}$$

where $\alpha_k = P(k|\{\beta_j^{(k)}\}_j)$ and an explicit dependence on these has been introduced into the notation on the LHS in the third line. In the second line we assume that the measurements of each object are independent, and in the third line conditional independence has been used to remove superfluous dependencies. The α_k parameters are just the same ones introduced in equation 4.

We can now maximize equation 5 (which we might call the *evidence*) w.r.t both $\{\beta_j^{(k)}\}$ and

$\{\alpha_k\}$, subject to the constraints

$$\begin{aligned}
\forall j,k: &\quad 0 \le \beta_j^{(k)} \le 1 \\
\forall k: &\quad \sum_j \beta_j^{(k)} = 1 \\
\forall k: &\quad 0 \le \alpha_k \le 1 \\
&\quad \sum_k \alpha_k = 1
\end{aligned} \tag{6}$$

2.1 Implementation

To maximize equation 5 w.r.t the normalization conditions in equation 6, it is easier to maximize its logarithm

$$\mathcal{L} = \sum_n \ln\left(\sum_k P(x_n|k, \{\beta_j^{(k)}\}_j, V_n)\,\alpha_k\right) \tag{7}$$

w.r.t the transformed variables $\{a_k\}$ and $\{b_j^{(k)}\}$, where

$$\begin{aligned}
\alpha_k &= \frac{\exp(a_k)}{\sum_{k'=1}^{K}\exp(a_{k'})} \\
\beta_j^{(k)} &= \frac{\exp(b_j^{(k)})}{\sum_{j'=1}^{J_k}\exp(b_{j'}^{(k)})} \quad \forall k\ .
\end{aligned} \tag{8}$$

This *softmax* transformation imposes the normalization conditions. These transformed variables are permitted to range between $\pm\infty$, and so are easier to work with in standard optimization algorithms. The optimization involves searching over a space of dimensionality $K + \sum_k J_k$, which is of the order of the number of templates, and is typically large. (This is one reason why we often choose to adopt continuous models instead.)

Note that equation 7 involves a sum over all templates (through equation 3) of the likelihood (equation 1), so every step in the optimizer has to calculate Gaussian exponential using all the data.

3 Semi-discrete models

The basic presentation of the previous section is somewhat restrictive because the templates are entirely fixed with no free parameters. Only their relevance in an inference is adjustable. But there are many applications in which it is useful to give the templates a variable “amplitude”. An example is when the data are photometric measurements of astronomical sources, i.e. x_n is a spectral energy distribution (SED). While we may consider a fixed set of templates, such as archetypal star or galaxy SEDs, these objects nonetheless appear at different apparent magnitudes in the universe. We don't want to have to duplicate all the templates at every possible apparent magnitude in order to accommodate this. We therefore generalize equation 1 to give each template a scalar amplitude, $c_j^{(k)}$

$$P(x_n|c_j^{(k)}, V_n, T_j^{(k)}, k) = \frac{1}{(2\pi)^{I/2}|V_n|^{1/2}} \exp\left[-\frac{1}{2}\left(x_n - c_j^{(k)}T_j^{(k)}\right)^{\mathsf{T}} V_n^{-1}\left(x_n - c_j^{(k)}T_j^{(k)}\right)\right] \tag{9}$$

The probability that the observed object is described by template $T_j^{(k)}$ is obtained by marginalizing over this amplitude

$$P(x_n|V_n, T_j^{(k)}, k) = \int_{c_j^{(k)}} P(x_n|c_j^{(k)}, V_n, T_j^{(k)}, k) P(c_j^{(k)}|T_j^{(k)}, k)\, dc_j^{(k)} \tag{10}$$

which is a one-dimensional integral. (Note that we can remove the V_n dependence from the second term.) This is just the above likelihood averaged over the prior for the amplitude. The rest of the inference proceeds as before from equation 2.

$P(c_j^{(k)}|T_j^{(k)}, k)$ is a one-dimensional function over $c_j^{(k)}$ for each template. Typically it must be positive (unless x_n can be negative), in which case a convenient choice might be the Gamma distribution. As this integral must be evaluated many times during the optimization, it is prudent to make it fast to calculate via some approximation. One possibility would be to write it as a weighted sum of fixed basis functions, fixed in the sense that they are functions only of the data (and so can be precalculated).

4 Marginalizing over the relevance parameters

If we are only interested in classifying the unlabelled data, then rather than maximizing $P(\{x_n\}|\{V_n\}, \{\beta_j^{(k)}\}, \alpha_k)$ w.r.t $\{\beta_j^{(k)}\}$ and $\{\alpha_k\}$ it is desirable to marginalize over these. We achieve this by extending the inference one level up by placing a prior probability distribution over $\{\beta_j^{(k)}\}$. ...

5 Summary

- In discrete classification, the fixed, class-labelled templates (perhaps with a variable amplitude) are used as they are. This is often referred to as *nonparametric* modelling. There is no forward model fitting.

- The relevances of the templates are inferred from the data using the whole data set including the unlabelled data. The model is therefore a type of *semi-supervised learning*.

- The model is hierarchical in the sense that both the template relevances (β) and the overall class probabilities (α) are inferred from the data.
- Learning or marginalizing over the model parameters involves integrations over a space with dimensionality of the order of the number of templates.