Exemplar-based likelihoods using the PDF projection theorem

Thomas Minka
Microsoft Research Cambridge
March 1, 2004

Abstract

In computer vision it is common to define algorithms in terms of matching against exemplars. This paper describes a probabilistic framework for such algorithms. It allows you to assign a consistent likelihood for each exemplar, eliminating the normalization and bias problems which occur under the scheme of Toyama & Blake (2002).

1 Introduction

Consider a vision task where we want to classify shapes by matching them to exemplars. To make the system modular and capable of handling prior knowledge, we want this process to be probabilistic. That is, we want a likelihood $p(x|c)$ which is high when the observation x matches the exemplar c, and low otherwise. We also want the likelihood to be normalized so that comparisons between $p(x|c=1)$ and $p(x|c=2)$ make sense.

Suppose our matching score is $f(x,c)$, e.g. the chamfer distance between x's edge map and the exemplar edge map. One common approach is to define

$$p(x|c) = \frac{1}{Z_c}\exp(\lambda_c f(x,c)) \quad (1)$$

$$\text{where } Z_c = \int_x \exp(\lambda_c f(x,c))dx \quad (2)$$

and subsequently to ignore the dependence of Z_c on c, thus dropping Z from the model (Toyama & Blake, 2002). Ignoring Z_c is equivalent to biasing the system toward exemplars with larger Z_c. These are the exemplars which give high match scores to random images, e.g. exemplars which are small or have very few edges. Such a system will prefer trivial matches to every image, unless you add some extra device to prevent it.

The correct way to use such a model is to actually compute the separate normalizers Z_c. However, for many matching scores of interest, such as Chamfer distance, this is intractable. The normalizer involves an integral over all possible observations, not just the ones we have observed in our training set. So another approach is needed.

2 The PDF projection theorem

A better approach is to use the PDF projection theorem (Baggenstoss, 2003) to define a tractable and properly normalized likelihood. The PDF projection theorem is a general method to convert a distribution on features to a distribution on raw data.

Let the raw data be x and the feature be $f = f(x, c)$, whose distribution we have modeled as $p(f|c)$. We want to find a distribution $p(x|c)$ such that when we apply the transformation f, we get exactly $p(f|c)$. This can be written as a linear equation:

$$p(f|c) = \int_x \delta(f - f(x, c))p(x|c)dx \tag{3}$$

(The f on the left refers to any value, while $f(x, c)$ refers to the specific value of f obtained from x.) In general, there are many solutions to this equation. We will pick the one which is closest to a given **base distribution** $g(x)$, which is a rough estimate of $p(x|c)$. This is analogous to picking the minimum-norm solution to a system of linear equations. Alternatively, you can think of it as *projecting* the base distribution onto the constraint set, which is a linear subspace. For the distance measure we will use KL-divergence, though the answer is the same for any divergence in a large family. Appendix A derives the solution:

$$p(x|c) = \frac{g(x)}{g(f)}p(f|c) \tag{4}$$

$$\text{where } g(f) = \int_x \delta(f - f(x, c))g(x)dx \tag{5}$$

The formula for $g(f)$ looks imposing, but it is simply the distribution of $f(x, c)$ when x's are drawn from $g(x)$. Because f is low-dimensional, this distribution is easy to model (just as easy as $p(f|c)$). The form in 4, while simple, is completely general in the sense that any $p(x|c)$ consistent with $p(f|c)$ can be written in this form.

Our exemplar-based model now works as follows. We have a training set of images x that originated from known exemplars c. For each exemplar c and images x generated from c, we model $f(x, c)$. A typical model might be exponential:

$$p(f(x, c)|c) = \lambda_c \exp(-\lambda_c f(x, c)) \tag{6}$$

Now we pick a base distribution, which for convenience can be the distribution of all the images in our training set. In other words,

$$g(x) = \sum_c p(x|c)p(c) \tag{7}$$

We model $f(x, c)$ for images from this distribution. This might also be exponential, or more generally a Gamma distribution. Call it $g(f(x, c))$. Then the likelihood for an exemplar is

$$p(x|c) = \frac{g(x)}{g(f(x, c))}p(f(x, c)|c) \tag{8}$$

and the posterior over exemplars given an image is

$$p(c|x) = \frac{p(c)p(x|c)}{\sum_c p(c)p(x|c)} = p(c)\frac{p(f(x, c)|c)}{g(f(x, c))} \tag{9}$$

This can be interpreted of normalizing the match score for each exemplar by its expected match score. However, it has a precise probabilistic meaning—notice (9) is an equality, not "proportional to".

Because it allows you to model images by only modeling features, the PDF projection theorem should have many applications throughout computer vision.

References

Baggenstoss, P. M. (2003). The PDF projection theorem and the class-specific method. *IEEE Transactions on Signal Processing*, *51*, 672–685.

Toyama, K., & Blake, A. (2002). Probabilistic tracking with exemplars in a metric space. *International Journal of Computer Vision*, *48*, 9–19. `http://research.microsoft.com/vision/cambridge/papers/toyama_ijcv02.pdf`.

Trottini, M., & Spezzaferri, F. (1999). A generalized predictive criterion for model selection (Technical Report 702). CMU Statistics Dept. `http://www.stat.cmu.edu/tr/`.

A Proof of the theorem

We want to minimize $KL(g(x) \,||\, p(x|c))$ subject to the constraints (3). Note that there is one constraint for every value of f. We don't need an explicit constraint for $\int_x p(x|c)dx = 1$ because the constraints require $\int_x p(x|c)dx = \int_f p(f|c)df = 1$.
Introducing Lagrange multipliers $\mu(f)$ gives the following objective:

$$\int_x g(x) \log \frac{g(x)}{p(x|c)} dx - \int_f \mu(f) \left(p(f|c) - \int_x \delta(f - f(x,c)) p(x|c) dx \right) df \tag{10}$$

Zeroing the derivative with respect to $p(x|c)$ at a particular x gives:

$$-\frac{g(x)}{p(x|c)} + \mu(f(x,c)) = 0 \tag{11}$$

$$p(x|c) = g(x)\mu(f(x,c)) \tag{12}$$

The constraints now require

$$p(f|c) = \int_x \delta(f - f(x,c)) g(x) \mu(f(x,c)) dx \tag{13}$$

$$= \mu(f) g(f) \tag{14}$$

Thus $\mu(f) = p(f|c)/g(f)$ and we obtain (4).
The above steps can be repeated for any divergence measure of the form

$$\frac{4}{1-\alpha^2} \int_x g(x) \left(1 - \frac{p(x|c)^{\frac{1+\alpha}{2}}}{g(x)^{\frac{1+\alpha}{2}}} \right) dx \tag{15}$$

which includes KL, reverse-KL, and Hellinger divergences (Trottini & Spezzaferri, 1999).