Hidden Variables

Qi Zhao

August 5, 2018

1. Introduction

To make the density multi-modal, it introduces mixture models [4]. To make the density robust, it replaces the normal with the t-distribution [2]. To cope with parameter estimation in high dimensions, it introduces subspace models [6].

The new models have much in common with each other. In each case it introduces a hidden or latent variable h_i associated with each observed data point x_i. The hidden variable induces the more complex properties of the resulting pdf. To model a complex probability density function over the variable x, it will introduce a hidden or latent variable h, which may be discrete or continuous.

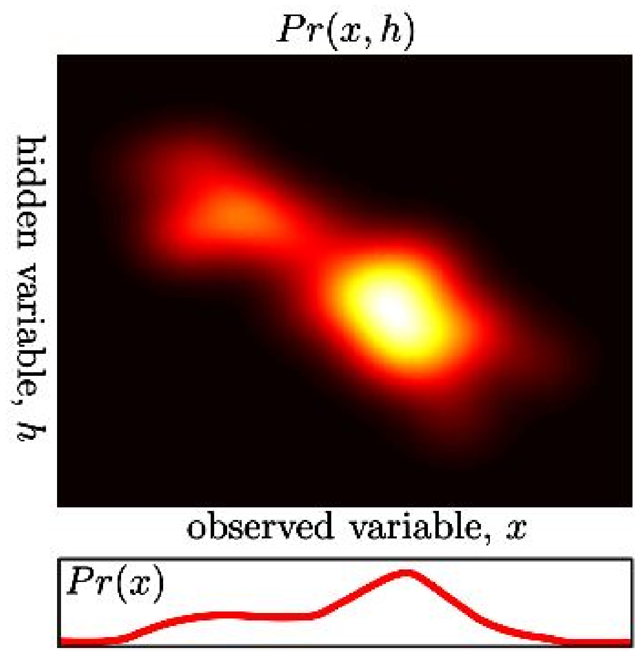

Figure 1. Using hidden variables to help model complex densities. One way to model the density Pr(x) is to consider the joint probability distribution Pr(x,h) between the observed data x and a hidden variable h. The density Pr(x) can be considered as the marginalization of (integral over) this distribution with respect to the hidden variable h. As we manipulate the parameters of this joint distribution, the marginal changes and the agreements with the observed data $\{x_i\}_{i=1}^{I}$ increases or decreases. Sometimes it is easier to fit the distribution in this indirect way than to directly manipulate Pr(x).

2. Descriptions

To exploit the hidden variables, it describes the final density Pr(x) as the marginalization of the joint density Pr(x,h) [5] between x and h in Equation 1 so that it is relatively simple to model, but produces an expressive family of marginal distributions Pr(x) when it integrates over h (see Figure 1).

$$Pr(x) = \int Pr(x,h)dh \tag{1}$$

Whatever form can be choose for the joint distribution in Equation 2, there are two possible approaches to fitting the model to training data $\{x_i\}_{i=1}^{I}$ using the maximum likelihood [3] method. It could directly maximize the log likelihood of the distribution Pr(x) from the left hand side of equation (see Equation 3). It uses the expectation maximization [1] algorithm, which works directly with the right-hand side of equation (see Equation 4).

$$Pr(x|\theta) = \int Pr(x,h|\theta)dh \tag{2}$$

$$\hat{\theta} = argmax_{\theta}[\sum_{i=1}^{I} logPr(x_i|\theta)] \tag{3}$$

$$\hat{\theta} = argmax_{\theta}[\sum_{i=1}^{I} log[\int Pr(x_i,h_i|\theta)dh_i]] \tag{4}$$

References

[1] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 24(8):1026–1038, 2002. 1

[2] E. A. Coornish. The multivariate t-distribution associated with a set of normal sample deviates. *Australian Journal of Physics*, 7(4):531–542, 1954. 1

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977. 1

[4] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 24(3):381–396, 2000. 1

[5] L. A. Shepp. The joint density of the maximum and its location for a wiener process with drift. *Journal of Applied Probability*, 16(2):423–427, 1979. 1

[6] M. Verhaegen and D. Westwick. Identifying MIMO wiener systems using subspace model identification methods. In *IEEE Conference on Decision and Control*, pages 4206–4211, 1995. 1