Fitting Probability Models [1]

Qi Zhao

July 16, 2018

1. Introduction

It concerns fitting probability models to data $x_i{}_i^I$ = 1. And It also concerns calculating the probability of a new datum x^* under the resulting model. This is known as evaluating the predictive distribution. We consider three methods: maximum likelihood [3], maximum a posteriori [4], and the Bayesian approach [2].

2. Conclusions

Maximum likelihood. As the name suggests, the maximum likelihood (ML) method finds the set of parameters $\hat{\theta}$ under which the data $x_i{}_i^I$ = 1 are most likely. To calculate the likelihood function Pr(x_i—θ) at a single data point x_i , we simply evaluate the probability density [5] function at x_i . Assuming each data point was drawn independently from the distribution, the likelihood function Pr($x_{1...I}$—θ) for a set of points is the product of the individual likelihoods. Hence, the ML estimate of the parameters is in Equation 1.

$$\begin{aligned} \hat{\theta} &= argmax_\theta[Pr(x_{1...I}|\theta)] \\ &= argmax_\theta \prod_{i=1}^{I} Pr(x_i|\theta) \end{aligned} \tag{1}$$

where $argmax_\theta$ f[θ] returns the value of θ that maximizes the argument f[θ].

Maximum a posteriori. In maximum a posteriori (MAP) fitting, From previous experience it may be known something about the possible parameter values. For example, in a time-sequence the values of the parameters at time t tell us a lot about the possible values at time t + 1, and this information would be encoded in the prior distribution. As the name suggests, maximum a posteriori in Equation 2 estimation maximizes the posterior probability Pr(θ—$x_{1...I}$) of the parameters. Comparing this to the maximum likelihood criterion, we can see that it is identical except for the additional prior term; maximum likelihood is a special case of maximum a posteriori where the prior is uninformative.

$$\hat{\theta} = argmax_\theta \prod_{i=1}^{I} [Pr(x_i|\theta)Pr(\theta)] \tag{2}$$

The Bayesian approach. Evaluating the predictive distribution is more difficult for the Bayesian case since it has not estimated a single model but have instead found a probability distribution over possible models. Hence, it can be calculated in Equation 3.

$$Pr(x^*|x_{1...I}) = \int Pr(x^*|\theta)Pr(\theta|x_{1...I})d\theta \tag{3}$$

which can be interpreted as follows: the term Pr(x^*—θ) is the prediction for a given value of θ. So, the integral can be thought of as a weighted sum of the predictions given by different parameters θ, where the weighting is determined by the posterior probability distribution Pr(θ—$x_{1...I}$) over the parameters (representing the confidence that different parameters are correct).

References

[1] S. M. Deni and A. A. Jemain. Fitting the distribution of dry and wet spells with alternative probability models. *Meteorology and Atmospheric Physics*, 104(1-2):13–27, 2009. 1

[2] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007. 1

[3] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981. 1

[4] J. L. Gauvain and C. H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994. 1

[5] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962. 1