Engineering and Applied Science Programs for Professionals
Whiting School of Engineering
Johns Hopkins University
685.621 Algorithms for Data Science
Machine Learning I

This document provides a rollup of the Machine Learning I. In this module the Expectation Maximization methods is introduced. This methods however is not a classifier on its own, A Bayes decision theory method is use to make a two-class classifiers.

Contents

1 Classification

Machine learning for a classification task involves training over a set of samples $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\lambda]$ where $\mathbf{x} \in \Upsilon_n$. Where the symbol Υ is used to represent the n-dimensional space the features in x resides. Each sample in the training set contains one target value $\mathbf{C} = C_j = [C_1, C_2, \cdots, C_c], j = 1, 2, \cdots, c$, (known as the class labels $y_i \in \mathbf{C}, i = 1, 2, \cdots, m$) which describes the class to which the sample is a member of. The objective is to separate the data into their classes such that the degree of association is strong between the data sets of the same class and weak between members of different classes. From the class separation, an unseen sample $\mathbf{x}_0 \in \Upsilon_n$ can then be appropriately classified. In this section 4 classification methods are presented, expectation maximization with mixture models (EM), k-nearest neighbors (k-NN), kernel Fisher's discriminant (KFD), and Parzen window.

1.1 Expectation Maximization

The idea behind the EM algorithm (Dempster et al., 1977) is that even though the data values of $\mathbf{x}$, feature vectors $\mathbf{x}_n \in \Upsilon_n$, are unknown/incomplete the distribution $f(\mathbf{x}|p)$ can be used to determine an estimate for the maximum likelihood (Tomasi, 2006). In maximum likelihood estimation, the estimate to be modeled is the parameter(s) for which the observed data are the most likely. This is done by iteratively estimating the data parameters, then using the data to update the estimated parameters, until a desired convergence is met. The two major steps of the EM algorithm are the expectation step (E-Step) and the maximization step (M-Step).

The EM algorithm consists of choosing initial parameters for the means, $\mu_k^{(j)}$, standard deviations, , and mixing probabilities, $\sigma_k^{(j)}$, for a user defined number of clusters, k, then performing the E-Step and M-Step successively until convergence, where i is the current iteration and n is the number of samples. The convergence criteria is determined by examining when the parameters quit changing, i.e., when $\left|\mu_k^{(j)} - \mu_k^{(j+1)}\right| < \epsilon$ & $\left|\sigma_k^{(j)} - \sigma_k^{(j+1)}\right| < \epsilon$ & $\left|p^{(j)}(k|l) - p^{(j+1)}(k|l)\right| < \epsilon$ for some epsilon (ϵ) and distance calculation (Euclidian distance). The maximum likelihood estimation is a method of estimating the parameters of the distributions based upon the observed data.

The expectation step (E-Step) calculates the membership probabilities, $p(k|l)$ (Tomasi, 2006). The mixing probabilities p_k are viewed as the sample mean of the membership probabilities $p(k|l)$ assuming a uniform distribution over all the data points. The Gaussian function, $g(\mathbf{x}; \mu_k^{(j)}, \sigma_k^{(j)})$, is used to compute mixture of Gaussian functions as shown in the denominator of $p(k|l)$.

$$p^{(j)}(k|l) = \frac{p_k^{(j)} g(\mathbf{x}; \mu_k^{(j)}, \sigma_k^{(j)})}{\sum_{k=1}^{K} p_k^{(j)} g(\mathbf{x}; \mu_k^{(j)}, \sigma_k^{(j)})} \tag{1}$$

$$g(\mathbf{x}; \mu_k^{(j)}, \sigma_k^{(j)}) = \frac{1}{\left(\sqrt{2\pi}\sigma_k\right)^n} \exp\left\{ -\frac{1}{2}\left(\frac{\|\mathbf{x} - \mu_k\|}{\sigma_k}\right)^2 \right\} \tag{2}$$

The maximization step (M-Step) uses the data from the expectation step as if it were measured data to determine the maximum likelihood estimate of the parameter (Tomasi, 2006). This estimated data is often referred to as the "imputed" data. This step is dependent upon the membership probabilities $p(k|l)$ which are computed in the E-Step. The EM algorithm consists of iterating the mean, standard deviation, and mixing probabilities until convergence. The mixing probabilities are the sample mean of the conditional probabilities $p(k|l)$ assuming a uniform distribution over all the data points.

$$\mu_k^{(j+1)} = \frac{\sum_{i=1}^{l} p^j(k|i)\mathbf{x}_i}{\sum_{i=1}^{l} p^j(k|i)} \tag{3}$$

$$\sigma_k^{(j+1)} = \sqrt{\frac{1}{D}\frac{\sum_{i=1}^{l} p^j(k|i)\left\|\mathbf{x}_i - \mu_k^{(j+1)}\right\|^2}{\sum_{i=1}^{l} p^j(k|l)}} \tag{4}$$

$$p_k^{(j+1)} = \frac{1}{l}\sum_{i=1}^{l} p^j(k|i) \tag{5}$$

1.1.1 Mixture Models

In mixture models, also known as model-based Gaussian clustering, the multivariate Gaussian normal is used as a density function similarly described in Equation 2. The general multivariate normal density for n dimensions is

$$g(\mathbf{x};\mu_k,\Sigma_k) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)\right\}}{\left(\sqrt{2\pi}\right)^n\left|\Sigma_k\right|^{1/2}}. \tag{6}$$

The geometric characteristics (size, shape and orientation) of the clusters are determined by the covariance matrix Σ_k which is generated in terms of eigenvalue decomposition described in Martinez and Martinez (2002). The decomposition of the covariance matrix Σ_k is used as a suitable model for the geometric characteristics of the cluster. The structure of the covariance matrix is as follows:

$$\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T \tag{7}$$

where λ_k is a scalar, D_k is the orthogonal matrix of eigenvectors and A_k is a diagonal matrix whose elements are proportional to the eigenvalues of Σ_k. Note that in EM the values p_k, μ_k, and σ_k are updated after each iteration and in the mixture models σ_k is replaced by Σ_k to represent the geometric characteristics of the clusters.

The eigenvalue decomposition can be modeled as various clustering arrangements. Celeux and Govaert (1995), describe in detail fourteen models based on the eigenvalue decomposition. Allowing for variations in the orientation, volume, shape and size of the clusters; six of these models are shown in Table 1 (Martinez and Martinez, 2002).

Table 1: Parameterization for Mixture Models

Model	Σ_k	**Geometric Shapel**	**Volume**	**Shape**	**Orientation**
1	$\lambda\mathbf{I}$	Spherical	Equal	Equal	N/A
2	$\lambda_k\mathbf{I}$	Spherical	Variable	Equal	N/A
3	$\lambda\mathbf{D}\mathbf{A}\mathbf{D}^T$	Ellipsoid	Equal	Equal	Equal
4	$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	Ellipsoid	Variable	Variable	Variable
5	$\lambda\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$	Ellipsoid	Equal	Equal	Variable
6	$\lambda_k\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$	Ellipsoid	Variable	Equal	Variable

The eigenvalue decomposition can be modeled as various clustering arrangements, i.e., spheres, ellipsoids and rotations of ellipsoids. Allowing the orientation, volume, shape and size of the clusters define the various models used. Figure 1 shows the mixture model using rotated ellipsoids (Model 4) to generate the decision boundary around each class.

1.1.2 Bayes Classifier

The EM algorithm can be used to find a class label for an input sample. Classification uses input samples described by feature vectors $\mathbf{x}_0 \in \Upsilon_n$ to assign the samples to a given class $\mathbf{C} = C_j = [C_1, C_2, \cdots, C_c], j = 1, 2, \cdots, c$.

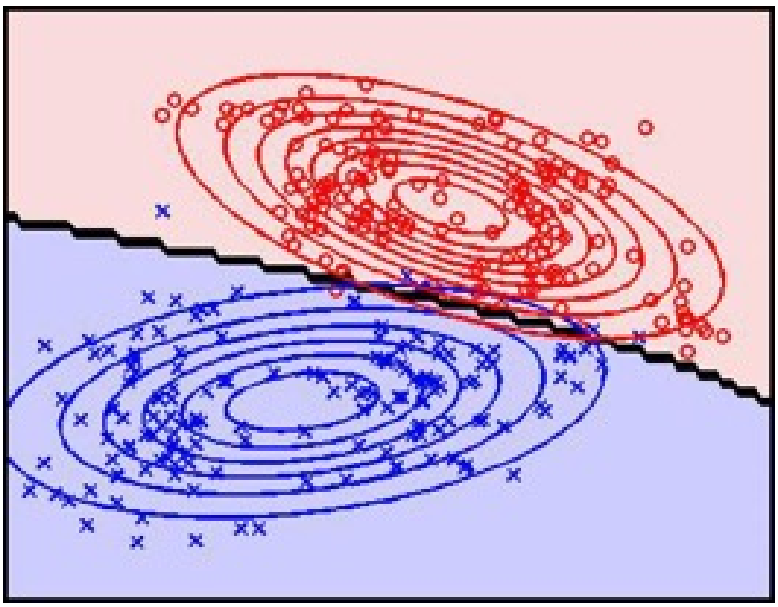

Figure 1: Expectation Maximization using mixture models with Decision Boundary

The Bayes classifier extends a general multivariate normal case where the covariance matrix Σ_j for each class is different. For the multi-class classifier each class must have individual conditional probability densities where the densities are modeled as normal distributions. The classes C_j are defined as normal distributions centered about the mean vector μ_j. The mean vector, μ_j, and the covariance matrix, Σ_j, are calculated using the EM algorithm. The vector $\mathbf{x}_0$ is a n-dimensional vector of the observed data, and $|\Sigma_i|$ and Σ_i^{-1} are the determinants and inverse covariance matrix of the given class. The posterior probability of class membership can be calculated by Bayes rule if C_j is defined as the event of belonging to population j. Using the density function $g(\mathbf{x}; \mu_k^{(i)}, \sigma_k^{(i)})$ (Tomasi, 2006), the Bayes classifier can be expressed in terms of the prior probabilities, $P(C_i)$, and posterior probability of class membership as follows:

$$P(C_j|x_0) = \frac{P(C_j)\frac{1}{\sqrt{(2\pi)^n|\Sigma_j|}}\exp\left[-\frac{1}{2}\left(\mathbf{x}_0-\mu_j\right)^T\Sigma_j^{-1}\left(\mathbf{x}_0-\mu_j\right)\right]}{\sum_{i=1}^{c}P(C_i)\frac{1}{\sqrt{(2\pi)^n|\Sigma_i|}}\exp\left[-\frac{1}{2}\left(\mathbf{x}_0-\mu_i\right)^T\Sigma_i^{-1}\left(\mathbf{x}_0-\mu_i\right)\right]} \tag{8}$$

where the a priori probabilities P(Cj) are the estimates of belonging to a class and under the assumption that $\Sigma_j = \Sigma$ for $\forall j$.

References

[1] Bishop, Christopher M., *Neural Networks for pattern Recognition*, Oxford University Press, 1995

[2] Bishop, Christopher M., *Pattern Recognition and Machine Learning*, Springer, 2006

[3] Cormen, Thomas H., Leiserson, Charles E., Rivest, Ronal L., and Stein, Clifford, *Introduction to Algorithms*, 3rd Edition, MIT Press, 2009

[4] Dempster, A. P., Laird, N. M. and Rubin, D. B., *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society B, Volume 39, Number 1, pp.1–22, 1977

[5] Duda, R.O., Hart, P.E. and Stork, D.G., *Pattern Classification* (2nd. Ed.), New York, NY: John Wiley & Sons, 2001

[6] Duin, Robert P.W., Tax, David and Pekalska, Elzbieta, *PRTools*, http://prtools.tudelft.nl/

[7] Franc, Vojtech and Hlavac, Vaclav, *Statistical Pattern Recognition Toolbox*, https://cmp.felk.cvut.cz/cmp/software/stprtool/index.html

[8] Fukunaga, Keinosuke, *Introduction to Statistical Pattern Recognition*, Academic Press, 1972

[9] Jaakola, T. S. and Haussler, D., *Exploring Generative Models in Discriminative Classifiers*, Advances in Neural Information Processing Systems, Kearns, M.S., Soll, S. A. and Cohn, D. A. (Eds.), Volume 11, Cambridge, MA: MIT Press, 1998

[10] Machine Learning at Waikato University, *WEKA*, https://www.cs.waikato.ac.nz/ ml/index.html

[11] Martinez W. L. and Martinez, A. R., *Computational Statistics Handbook with MATLAB*, Boca Raton, FL: Chapman & Hall/CRC, 2002

[12] Parzen, E., *On the Estimation of a Probability Density Function and Mode*, Annals of Mathematical Statistics, Volume 33 pp. 1065-1076, 1962

[13] Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P., *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Jan 31, 1986

[14] Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P., *Numerical Recipes: The Art of Scientific Computing*, 3rd Edition, Cambridge University Press, September 10, 2007

[15] Press, William H., Teukolsky, Saul A., Vetterling, William T., and Flannery, Brian P., *Numerical Recipes: The Art of Scientific Computing*, 3rd Edition, http://numerical.recipes/

[16] Press, William H., *Opinionated Lessons in Statistics*, http://www.opinionatedlessons.org/

[17] Tomasi, C., *Estimating Gaussian Mixture Densities with EM – A Tutorial*, Duke University Course Notes, 2006, http://www.cs.duke.edu/courses/spring04/cps196.1/handouts/EM/tomasiEM.pdf, Retrieved Sept 2006