The Abdus Salam
International Centre
for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

2361-18

School on Large Scale Problems in Machine Learning and Workshop on Common Concepts in Machine Learning and Statistical Physics

20 - 31 August 2012

MACHINE LEARNING IN SYSTEMS BIOLOGY: Probabilistic Classification for Large Dimensionality - EXERCISES

Ole WINTHER
Technical University of Denmark DTU and University of Copenhagen KU
Denmark

Probabilistic classification for large dimensionality

August 21, 2012

1 Exercises for lectures

The Exercises will solved during the lectures. We will focus on the probabilistic generative models for classification (also know as the Bayes classifier) because this is a simple model of practical interest which can be analysed in some detail. In addition to these pen and paper exercises, those interested may also work with hands-on-data Matlab factor modeling exercises which may be retrieved here `http://www.imm.dtu.dk/Forskning/ISP/Undervisning/02901_2012.aspx` .

Exercise 1 - Derive Bayes classifier

The Bayes classifier (or probabilitistic generative approach, see C Bishop, Pattern Recognition and Machine Learning, Section 4.2) maps an input (co-variate) $\mathbf{x}$ to the a probability for a class $\mathcal{C}_k$, $k = 1, \ldots, K$ using Bayes' theorem:

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} \qquad \text{with} \qquad p(\mathbf{x}) = \sum_{k'=1}^{K} p(\mathbf{x}|\mathcal{C}_{k'})p(\mathcal{C}_{k'}) \ .$$

So given the density of the covariates for each class and the prior probabilities of the classes we can compute class posterior probabilities. We can also write this in terms of the so-called soft-max function:

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{\exp(a_k)}{\sum_{k'=1}^{K} \exp(a_{k'})}$$

with $a_k = \log p(\mathbf{x}|\mathcal{C}_k) + \log p(\mathcal{C}_k)$. We will choose Gaussian class-conditional densities:

$$p(\mathbf{x}|\mathcal{C}_k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \equiv \frac{1}{\sqrt{\det 2\pi\boldsymbol{\Sigma}_k}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x}-\boldsymbol{\mu}_k)\right\} \ .$$

Questions:

1. Show that for $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ we have
$$a_k = \mathbf{w}_k^T\mathbf{x} + w_{k0}$$
with $\mathbf{w}_k = \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k$ and $w_{k0} = -\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k/2 + \log p(\mathcal{C}_k)$.

2. Consider two classes $K = 2$. (Results can easily be generalized.) The decision boundary is points in $\mathbf{x}$-space where $p(\mathcal{C}_1|\mathbf{x}) = p(\mathcal{C}_2|\mathbf{x}) = 0.5$. Show that this condition leads to a linear equation in $\mathbf{x}$ and thus a linear decision boundary. The model is also known as linear discriminant analysis.

3. We now allow different covariances $\boldsymbol{\Sigma}_k$, $k = 1, \ldots, K$. Show that the decision boundaries now become quadratic in $\mathbf{x}$. This model is also know as quadratic discriminant analysis.

Exercise 2 - Maximum likelihood estimation

We will now use maximum likelihood to estimate the parameters $\theta \equiv \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)|k = 1, \ldots, K\}$ from a training set of input-output pairs $(\mathbf{X}, \mathbf{T}) = \{(\mathbf{x}_n, \mathbf{t}_n)|n = 1, \ldots, N\}$, where $\pi_k = p(\mathcal{C}_k)$ and t_{nk} is an indicator variable being one if example n belongs to class $\mathcal{C}_k$ and zero otherwise. The model is generative so the likelihood is the joint probability of input and output pairs which we decompose into the class prior and class-conditional: $p(\mathbf{x}_n, \mathcal{C}_k|\theta) = p(\mathcal{C}_k|\theta)p(x_n|\mathcal{C}_k, \theta)$. Assuming identically independently distributed data the likelihood for the parameters $p(\mathbf{X}, \mathbf{T}|\theta) = prod_n p(\mathbf{x}_n, \mathbf{t}_n|\theta)$ with shorthand $\theta = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)|k = 1, \ldots, K\}$ we have

$$p(\mathbf{x}_n, \mathbf{t}_n|\theta) = \prod_k \left[p(\mathcal{C}_k)p(\mathbf{x}_n|\mathcal{C}_k)\right]^{t_{nk}} = \prod_k \left[\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right]^{t_{nk}} \ .$$

Note that the indicator variable will select the term in the product corresponding to the class from which the example is taken.

Questions:

1. Derive the maximum likelihood solution

$$\begin{aligned}
\pi_{k,\mathrm{ML}} &= \frac{N_k}{N} \qquad \text{with} \qquad N_k = \sum_{n=1}^{N} t_{nk} \\
\boldsymbol{\mu}_{k,\mathrm{ML}} &= \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} \mathbf{x}_n \\
\boldsymbol{\Sigma}_{k,\mathrm{ML}} &= \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_{k,\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{k,\mathrm{ML}})^T
\end{aligned}$$

   by setting the derivative of the log likelihood equal to zero and handling the sum to one constraint $\sum_k \pi_k = 1$ by a Lagrange multiplier. (If uncomfortable with Lagrange multipliers then consider $K = 2$ and $\pi_2 = 1 - \pi_1$.) This result simply stated says that the maximum liklihood estimate coincides with the empirical estimate of the three quantities

2. Show that if we again set $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ we get a maximum likelihood estimate as a weighted average of the covariances for each class

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N} \sum_{k=1}^{K} N_k \boldsymbol{\Sigma}_{k,\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_{k,\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{k,\mathrm{ML}})^T \ .$$

Exercise 3 - Spectral decomposition

In order to understand why the maximum likelihood estimate of the covariance matrix in general will not be robust we can make an eigenvalue decomposition of the empirical covariance matrix for examples belonging to class k: $\boldsymbol{\Sigma}_{k,\mathrm{ML}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ where the columns of $\mathbf{U}$ are the eigenvector of $\boldsymbol{\Sigma}_{k,\mathrm{ML}}$ and $\boldsymbol{\Lambda}$ is diagonal with the non-negative eigenvalues on the diagonal. At most $\min(d, N_k - 1)$ eigenvalues are non-zero. The minus one comes the fact that we have used 'one degree of freedom' to estimate $\boldsymbol{\mu}_k$. For the shared covariance case at most $\min(d, N - K)$ eigenvalues are non-zero.

Questions:

1. we will now consider the shared covariance case. Show that

$$a_k = \mathbf{w}_k^T \mathbf{x} + w_{k0} = \sum_{i=1}^{d} \frac{\alpha_{ik}}{\lambda_i} \mathbf{u}_i^T \mathbf{x} + w_{k0} \ ,$$

   where $\alpha_{ik} = \mathbf{u}_i^T \boldsymbol{\mu}_k$ is the projection of the mean on eigenvector i and λ_i is the corresponding eigenvalue.

2. Consider now $N < d$ as in the tumor gene expression classification task. Why is it not possible to use the maximum likelihood solution as it stands? Can you propose a simple modification to avoid getting a contribution from the directions with no empirical variance?

Exercise 3 - Calibrating the classifier for $N < d$

A second related problem of working with high dimensional data is that the probabilistic classifier is typically not well-calibrated. In a well-calibrated classifier the posterior class probabilities should coincide with the actual probability of the different outcomes. For the tumor classification task we observe posterior class probabilities that tend to be quite extreme, that is the class with highest probability is very close to one. In practice, clinicians expect the task to be difficult and this is also what our test and cross-validation results show. One source of error is that $p(\mathbf{x}|\mathcal{C}_k)$ is definitely not Gaussian and one might also do better with alternative discriminative approaches (modeling $p(\mathcal{C}_k|\mathbf{x})$ directly), however the problem of calibration is probably a quite characteristic feature of working with $N < d$ and high d.

Questions:

1. We can get some idea about why this happens by inspecting the expression for the posterior probabilities. Here we consider binary classification, $K = 2$. Show that

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{1}{1+\exp(-\Delta a)} \qquad \text{with} \qquad \Delta a = \sum_{i=1}^{d} \frac{\Delta\alpha_i}{\lambda_i}\mathbf{u}_i^T\mathbf{x} + \Delta w_0 \ .$$

   Express Δa in terms of α and λ using the basis expansion result: $\boldsymbol{\mu}_k = \sum_i \alpha_{ki}\mathbf{u}_i$. Derive the expression for Δa when $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are expanded by disjoint basis vectors. can we say something about the order of Δa as a function of d assuming for simplicity that $\mathbf{x}^T\mathbf{x} = \boldsymbol{\mu}_k^T\boldsymbol{\mu}_k = 1$?

2. This paper `http://www.cs.toronto.edu/~radford/selbias-ba.abstract.html` discusses another aspect of working with high dimensional data and feature selection. Describe the general principle without going into technical details.

Exercise 4 - Structured covariance estimation

The maximum likelihood solution in pPCA for $\mathbf{W}$ for M factors (Bishop Section 12.2.1) is

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathrm{ML}} &= \mathbf{W}_{\mathrm{ML}}\mathbf{W}_{\mathrm{ML}}^T + \sigma_{\mathrm{ML}}^2\mathbf{I} \\
\mathbf{W}_{\mathrm{ML}} &= \mathbf{U}_M(\boldsymbol{\Lambda}_M - \sigma_{\mathrm{ML}}^2\mathbf{I}_M)^{1/2}\mathbf{R}_M \\
\sigma_{\mathrm{ML}}^2 &= \frac{1}{d-M}\sum_{i=m+1}^{d}\lambda_i \ ,
\end{aligned}$$

where $\boldsymbol{\Lambda}_M$ is submatrix of $\boldsymbol{\Lambda}$ with the M largest eigenvalues and $\mathbf{R}$ is an arbitrary unitary (rotation) matrix, that is $\mathbf{R}\mathbf{R}^T = \mathbf{I}$. Note that σ_{ML}^2 is average of the variance in principal directions not captured by $\mathbf{W}_{\mathrm{ML}}$.

Questions:

1. Insert the pPCA result and calculate a_k for this case. Will this model have a better behavior for $N < d$?

2. Discuss ways of selecting M, the number of factors.

Ole Winther, August 2012.