Assignment #1 – AOS1

Bayesian inference

This homework aims at programming and testing regularized Gaussian discriminant analysis on several datasets, and to compare it with logistic regression.

You will have to provide a report (LaTeX recommended, 10 pages maximum) with your code *in an appendix*, which will provide a presentation and some comments on the results obtained.

You will also provide the Python code with your functions and the script with your training/test procedure separately, so that it can be executed.

The projects are due for Nov. 2nd, 2020.

1 Discriminant analysis

1.1 Model

We consider data in the form of feature vectors $\boldsymbol{x} \in \mathbb{R}^d$ associated with class information $z \in \Omega = \{\omega_1, \ldots, \omega_K\}$. We want to classify any (new) instance $\boldsymbol{x}$, for which z is unknown, based on a distribution of posterior probabilities over the classes computed from $\boldsymbol{x}$. In Gaussian discriminant analysis, we use the following model for this purpose. For any $k = 1, \ldots, K$,

$$\Pr(\omega_k|\boldsymbol{x};\theta) = \frac{\pi_k f_k(\boldsymbol{x};\theta_k)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\boldsymbol{x};\theta_\ell)}, \tag{1}$$

here, $\theta = \{\theta_1, \ldots, \theta_K\}$ is the set of all parameters to estimate, and $\theta_k = (\pi_k, \mu_k, \Sigma_k)$ is the set of parameters for class ω_k:

- the class proportion $\pi_k \in [0; 1]$,
- the expectation vector $\mu_k \in \mathbb{R}^d$,
- the covariance matrix $\Sigma_k \in \mathcal{M}_{d,d}$ (which is in addition symmetric, positive-definite).

The class-conditional density $f_k(\boldsymbol{x};\theta_k) = \Pr(\boldsymbol{X} = \boldsymbol{x}|Z = \omega_k)$ (i.e., the density at $\boldsymbol{x}$ in class ω_k) is that of the multivariate Gaussian distribution:

$$f_k(\boldsymbol{x}) = (2\pi)^{-\frac{d}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \mu_k)^\top \Sigma_k^{-1} (\boldsymbol{x} - \mu_k)\right).$$

In a nutshell, this model assumes that the distribution of the instances in each class is (multivariate) Gaussian. The shape of the Gaussian distribution (orientation and volume) in class ω_k is defined by the covariance matrix Σ_k; its location in the input space, by the expectation vector μ_k; and its relative weight, by the class proportion π_k. All of these parameters are generally unknown, and must therefore be estimated from data.

1.2 Training

We have a training set of labeled instances $\{(\boldsymbol{x}_1, z_1), \ldots, (\boldsymbol{x}_n, z_n)\}$. We will assume that the class information for instance $\boldsymbol{x}_i$ is encoded by a vector $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$, with $z_{ik} = 1$ if $\boldsymbol{x}_i \in \omega_k$ and $z_{ik} = 0$ otherwise.

During training, the model parameters, i.e. π_k, μ_k and Σ_k for each class ω_k, $k = 1, \ldots, K$, are generally estimated by maximum likelihood. The parameter estimates are detailed in Appendix A.

1.3 Test

Once the parameters are estimated, any new instance can be classified by computing the posterior probabilities according to Equation (1), where the parameters π_k, μ_k and Σ_k are replaced by their estimates; and by choosing the class with maximal posterior probability. It can be shown that the corresponding decision boundaries are quadratic.

Additional assumptions on the covariance matrices result in a simplified model, with fewer parameters to be estimated — which should thus be more robust to the lack of data. For instance, each matrix Σ_k can be assumed to be diagonal; or all matrices can be assumed to be identical, i.e. $\Sigma_1 = \cdots = \Sigma_K$.

1.4 Prior choice

A classical choice consists in setting a Gaussian-inverse-Wishart prior on the class expectations and covariance matrices (the class proportions can be handled separately using a Dirichlet prior, which you are not required to consider here). In a nutshell, a Gaussian prior is put on each expectation vector, which depends on the covariance matrix Σ_k:

$$\begin{aligned}\mu_k|\Sigma_k &\sim \mathcal{N}\left(\mu_{kp}, \Sigma_k/\kappa_{kp}\right), \\ &\propto |\Sigma_k|^{-1/2} \exp\left(-\frac{\kappa_{kp}}{2}(\mu_k - \mu_{kp})^\top \Sigma_k^{-1}(\mu_k - \mu_{kp})\right);\end{aligned}$$

and an inverse-Wishart prior is put the associated covariance matrix:

$$\begin{aligned}\Sigma_k &\sim \text{invW}\left(\Lambda_{kp}, \nu_{kp}\right), \\ &\propto |\Sigma_k|^{-\frac{\nu_{kp}+d+1}{2}} \exp\left(-\frac{1}{2}\,\text{trace}\left(\Sigma_k^{-1}\Lambda_{kp}^{-1}\right)\right).\end{aligned}$$

This prior distribution and the multivariate Gaussian sampling distribution are conjugate, and thus using this prior simply amounts to updating the parameter estimates according to the prior hyper-parameters.

The hyper-parameters are generally chosen by the user. The resulting penalized estimates are provided in Appendix A.

2 Questions

2.1 Classical discriminant analysis

(1) Program Gaussian discriminant analysis. You will make two functions. The first one will train the model based on a training set (design matrix `Xtr` and partition matrix `ztr`). You will only consider the general model (i.e., without any additional assumption on the covariance matrices). The second function will take a set of instances `Xte` and provide the corresponding posterior probabilities and associated decisions.

It is recommended that you use the appropriate functions for computing the empirical covariance and the empirical mean of a set of vectors stored into a design matrix. As well, the functions related to the multivariate Gaussian distribution (and in particular the one which computes the multivariate Gaussian density for some data and given some parameters) will be of a great use.

(2) Test your model on the datasets provided:

- the synthetic dataset in `synth`, where the classes are Gaussian;
- the real datasets in `bcw` (which stands for "breast cancer Wisconsin"), `Pima`, and `spambase`.

Compare the results with those obtained with quadratic logistic regression. What do you conclude out of these experiments ?

2.2 Regularized discriminant analysis

(3) Program the regularized version of Gaussian discriminant analysis. Once classical Gaussian DA is programmed, you just have to add new parameters to the function which are the hyper-parameters corresponding to the prior distributions. Only the function dedicated to training needs to be adapted.

(4) Test the regularized version of your model on the datasets provided (`synth`, `bcw`, `Pima`, and `spambase`). Compare the results with those obtained with penalized quadratic logistic regression. Again, what do you conclude ?

A Parameter estimates

A.1 Maximum likelihood estimates (non-regularized model)

Class proportions The class proportions are classically estimated by

$$\widehat{\pi}_k = \frac{\sum_{i=1}^{n} z_{ik}}{n} = \frac{n_k}{n},$$

that is, the empirical proportion of instances from class ω_k. Note that here, $n_k = \sum z_{ik}$ is the number of training instances labeled as being from class ω_k.

Expectation vectors The expectation vectors are classically estimated by

$$\widehat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik} \boldsymbol{x}_i = \overline{\boldsymbol{x}}_k,$$

that is, the empirical average of the training instances labeled as being from class ω_k.

Covariance matrices The covariance matrices are classically estimated by

$$\widehat{\Sigma}_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik} (\boldsymbol{x}_i - \widehat{\mu}_k)(\boldsymbol{x}_i - \widehat{\mu}_k)^\top = V_k,$$

which is the empirical covariance matrix of the training instances labeled as being from class ω_k.

A.2 Maximum a posteriori estimates (regularized model)

Gaussian inverse-Wishart prior We consider here a Gaussian-inverse-Wishart prior set on the parameters (μ_k, Σ_k) for class ω_k:

$$\begin{aligned}
\mu_k | \Sigma_k &\sim \mathcal{N}\left(\mu_{kp}, \Sigma_k / \kappa_{kp}\right), \\
&\propto |\Sigma_k|^{-1/2} \exp\left(-\frac{\kappa_{kp}}{2} (\mu_k - \mu_{kp})^\top \Sigma_k^{-1} (\mu_k - \mu_{kp})\right); \\
\Sigma_k &\sim \text{invW}\left(\Lambda_{kp}, \nu_{kp}\right), \\
&\propto |\Sigma_k|^{-\frac{\nu_{kp}+d+1}{2}} \exp\left(-\frac{1}{2} \text{trace}\left(\Sigma_k^{-1} \Lambda_{kp}^{-1}\right)\right).
\end{aligned}$$

The hyper-parameter μ_{kp} is obviously the mean of the Gaussian prior distribution set on μ_k, and therefore defines the expected value (and its mode); besides, κ_{kp} is called the shrinkage parameter for the Gaussian prior: the higher its value, the "more concentrated" the prior will be around its mean. The hyper-parameter Λ_{kp} is called the scale of the inverse-Wishart prior, and influences the mean (and the mode) of the distribution; again, ν_{kp} represents the number of degrees of freedom of the distribution.

In this case, the maximum-likelihood estimates given above can be replaced by their maximum a posteriori counterparts derived from the posterior distribution:

$$\begin{aligned}
\widehat{\mu}_k &= \frac{n_k \overline{\boldsymbol{x}}_k + \kappa_{kp} \mu_{kp}}{n_k + \kappa_{kp}}, \\
\widehat{\Sigma}_k &= \frac{n_k V_k + \dfrac{n_k \kappa_{kp}}{n_k + \kappa_{kp}} (\overline{\boldsymbol{x}}_k - \mu_{kp}) (\overline{\boldsymbol{x}}_k - \mu_{kp})^\top + \Lambda_{kp}^{-1}}{n_k + \nu_{kp} + d + 2}.
\end{aligned}$$