IMPACT: International Journal of Research in Applied, Natural and Social Sciences (IMPACT: IJRANSS)
ISSN(E): 2321-8851; ISSN(P): 2347-4580
Vol. 3, Issue 8, Aug 2015, 57-70

NON-LINEAR KERNEL FISHER DISCRIMINANT ANALYSIS WITH APPLICATION

CHRO K. SALIH

Lecture**,** Department of Mathematics, Faculty of Science and Education Science, School of Education Science, Sulaimania University, Iraq

ABSTRACT

Linear Discriminant Analysis (LDA) is a traditional statistical method which has proven successful on classification and dimensionality reduction problems [(5)]. The procedure is based on an eigenvalue resolution and gives an exact solution of the maximum of the inertia but this method fails for a nonlinear problem.

To solve this problem used kernel Fisher Discriminant analysis (KFDA), carry out Fisher linear Discriminant analysis in a high dimensional feature space defined implicitly by a kernel. The performance of KFDA depends on the choice of the kernel.

In this paper, we consider the problem of finding the optimal solution over a given linear kernel function for the two primal and dual variable in Fisher Discriminant, this by taking a small sample 20 case about HIV disease by taking three factors (Age, Gender, number of Lymphocyte cell) with two level to clear how these observations classified by testing this classified using statistic (Rayleigh Coefficient).

KEYWORDS: Linear Fisher Discriminant, Kernel Fisher Discriminant, Rayleigh Coefficient, Cross-Validation, Regularization

INTRODUCTION

Fisher's linear Discriminant separates classes of data by selecting the features that maximize the ratio of projected class means to projected intraclass variances. [(3)]

The intuition behind Fisher's linear Discriminant (FLD) consists of looking for a vector of compounds $\mathbf{w}$ such that, when a set of training samples are projected into it, the class centers are far apart while the spread within each class is small, consequently producing a small overlap between classes[(11)]. This is done by maximizing a cost function known in some contexts as Rayleigh Coefficient, $J(\mathbf{w})$.

Kernel Fisher's Discriminant (KFD) is a nonlinear station that follows the same principle for Fisher Linear Discriminant but in a typically high-dimensional feature space $\mathcal{F}$. In this case, the algorithm is reformulated in terms of $J(\alpha)$, where α is the new direction of Discriminant. The theory of reproducing kernels in Hilbert space[(1)] gives the relation between vectors $\mathbf{w}$ *and* α. In either case, the objective is to determine the most "plausible" direction according to the statistic J.[(8)] demonstrated that KFD can be applied to classification problems with competitive results. KFD shares many of the virtues of other kernel based algorithms: the appealing interpretation of a kernel as a mapping of an input to a

high dimensional space and good performance in real life applications, among, the most important. However, it also suffer from the deficiencies of kernelized algorithms: the solution will typically include a regularization coefficient to limit model complexity and parameter estimation will rely on some from while the latter precludes the use of richer models.

Recently, KFDA has received a lot of interest in the literature[14, 9]. A main advantage of KFDA over other kernel-based methods is that computationally simple: it requires the factorization of the Gram matrix computed with given training examples, unlike other methods which solve dense (convex) optimization problems.

THEORETICAL PART

Notation and Definitions

We use $\mathcal{X}$ to denote the input or instance set, which is an arbitrary subset of $\mathbb{R}^n$, and $\mathcal{Y} = \{-1,+1\}$ to denote the output or class label set. An input-output pair (x, y) where $x \in \mathcal{X}$ and $\mathcal{Y} \in \mathcal{Y}$ is called an example. An example is called positive or (negative) if its class label is $+1(-1)$. We assume that the examples are drawn randomly and independently from a fixed, but unknown, probability distribution over $\mathcal{X} \times \mathcal{Y}$.

Asymmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel (function) if it is satisfies the finitely positive semi-definite property: for any $x_1, x_2, ..., x_m \in \mathcal{X}$, the Gram matrix $G \in \mathbb{R}^{m \times m}$, defined by

$$G_{ij} = K(x_i, x_j) \tag{1}$$

is positive semi-definite. Mercer's theorem [12] tells us that any kernel function K implicitly maps the input set $\mathcal{X}$ to a high dimensional (possibly infinite) Hilbert space $\mathcal{H}$ equipped with the inner product $\langle \cdot,\cdot \rangle_{\mathcal{H}}$ through a mapping $\phi : \mathcal{X} \to \mathcal{H}$:

$$K(x, z) = \langle \phi(x), \phi(z), \rangle_{\mathcal{H}}, \forall x, z \in \mathcal{X}$$

We often write the inner product $\langle \phi(x), \phi(z), \rangle_{\mathcal{H}}$ as $\phi(x)^T \phi(z)$, when the Hilbert space is clear from the context. This space is called the *feature space*, and the mapping is called the *feature mapping*. The depend on the kernel function K and will be denoted as ϕ_K *and* $\mathcal{H}_K$. The gram matrix $G \in \mathbb{R}^{m \times m}$ defined in (1), will be denoted G_K when it is necessary to indicate the dependence on.[7]

FISHER DISCRIMINANT

Fisher Discriminant is the earliest approaches to the problem of classification learning. The idea underlying this approach is slightly different from the ideas outlined so far, rather than using decomposition $\mathbf{P}_{xy} = \mathbf{P}_{y|x}\mathbf{P}_x$ we now decompose the unknown probability measure constituting the learning problem as $\mathbf{P}_{xy} = \mathbf{P}_{x|y}\mathbf{P}_y$. The essential different between these two formal expression becomes apparent when considering the model choices :

- In the case of $\mathbf{P}_{xy} = \mathbf{P}_{y|x}\mathbf{P}_x$ we use hypotheses $h \in \mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ to model the conditional measure $\mathbf{P}_{y|x}$ of

classes $y \in \mathcal{Y}$ given objects $x \in \mathcal{X}$ and marginalize over $\mathbf{P}_{\mathbf{x}}$ in the noise free case, each hypothesis defines such a model by $P_{\mathsf{Y}|\mathsf{X}=x,\mathbf{H}=h}(y) = \mathbf{I}_{h(\mathbf{x})=y}$. Since our model for learning contains only predictors $h : \mathcal{X} \to \mathcal{Y}$ that discriminate between objects, this approach is sometimes called the predictive or discriminative approach.

- In the case of $\mathbf{P}_{\mathbf{xy}} = \mathbf{P}_{\mathbf{x}|\mathbf{y}}\mathbf{P}_{\mathbf{y}}$ we model the generation of objects $x \in \mathcal{X}$ given the class $y \in \mathcal{Y} = \{-1,+1\}$ by some assumed probability model $\mathbf{P}_{\mathbf{x}|\mathbf{y}=y,\mathbf{Q}=\theta}$ where $(\theta_{+1}, \theta_{-1}, P) \in \mathrm{Q}$ parameterizes this generation process. We have the additional parameter $P \in [0,1]$ to describe the probability $\mathbf{P}_{\mathsf{Y}|\mathsf{Q}=\theta}(y)$ by $p \cdot \mathbf{I}_{y=+1} + (1-p)\mathbf{I}_{y=-1}$. As the model Q contains probability measures from which the generated training sample $x \in \mathcal{X}$ is sampled, this approach is sometimes called the generative or sampling approach.

In order to classify a new test object $x \in \mathcal{X}$ with a model in the generative approach we make use of Bayes theorem, i.e

$$\mathbf{P}_{\mathsf{Y}|\mathsf{X}=x,\mathsf{Q}=\theta}(y) = \frac{\mathbf{P}_{\mathsf{Y}|\mathsf{X}=x,\mathsf{Q}=\theta}(x)\mathbf{P}_{\mathsf{Y}|\mathsf{Q}=\theta}(y)}{\sum_{\tilde{y}\in\mathcal{Y}} \mathbf{P}_{\mathsf{X}|\mathsf{Y}=\tilde{y},\mathsf{Q}=\theta}(x)\mathbf{P}_{\mathsf{Y}|\mathsf{Q}=\theta}(\tilde{y})}.$$

In the case of two classes and the zero-one loss $l_{0-1}(h(x), y) \overset{def}{=} \mathbf{I}_{h(x)\neq y}$, we obtain for Bayes optimal classification at a novel test object $x \in \mathcal{X}$,

$$h_\theta(x) = \arg\max_{y=\{-1,+1\}} \mathbf{P}_{\mathsf{Y}|\mathsf{X}=x}(y)$$

$$= sign\left(\ln\left(\frac{\mathbf{P}_{\mathsf{X}|\mathsf{Y}=+1,\mathsf{Q}=\theta}(x).p}{\mathbf{P}_{\mathsf{X}|\mathsf{Y}=-1,\mathsf{Q}=\theta}(x)(1-p)} \right) \right) \quad (2)$$

as the fraction of this expression is greater than one if, and only if, $\mathbf{P}_{\mathsf{XY}|\mathsf{Q}=\theta}(x,+1)$ is greater than $\mathbf{P}_{\mathsf{XY}|\mathsf{Q}=\theta}(x,-1)$ in the generative approach the task of learning amounts to finding the parameters $\theta^* \in \mathrm{Q}$ or measures $\mathbf{P}_{\mathsf{X}|\mathsf{Y}=y,\mathsf{Q}=\theta^*}$ and $\mathbf{P}_{\mathsf{Y}|\mathsf{Q}=\theta^*}$ which incur the smallest expected risk $R(h_{\theta^*})$ by virtue of equation (2). Again, we are faced with the problem that, with out restrictions on the measure $\mathbf{P}_{\mathsf{X}|\mathsf{Y}=y}$, the best model is the empirical measure $\mathbf{v}_{\mathrm{x}_y}(x)$[1] where $x_y \subseteq x$ is the sample of all training objects of class y. Obviously, this is a bad model because $\mathbf{v}_{\mathrm{x}_y}(x)$ as-signs zero probability to all test objects not present in the training sample and thus $h_\theta(x) = 0$, i.e. we are unable to make predictions on unseen objects. Similarly to the choice of the hypothesis space in the Discriminative model we most

constrain the possible generative models $\mathbf{P}_{X|Y=y}$.

Let us consider the class of probability measures from the exponential family

$$\mathbf{P}_{X|Y=y,Q=\theta}(x)=a_0(\theta_y)\tau_0(x)\exp(\theta'_y(\tau(x)))$$

For some fixed function $a_0:Q\to\mathbb{R}$, $\tau_0:\mathcal{X}\to\mathbb{R}$ and $\tau:\mathcal{X}\to K$ using this functional form of the density we see that each decision function h_θ must be of the following form

$$h_\theta(x)=sign\left(\ln\left(\frac{a_0(\theta_{+1})\tau_0(x)\exp(\theta'_{+1}(\tau(x))).p}{a_0(\theta_{-1})\tau_0(x)\exp(\theta'_{-1}(\tau(x))).(1-p)}\right)\right)$$

${}^1\mathbf{v}_{x_y}(x)$ empirical probability measure

$$=sign+\ln\left(\underbrace{(\theta_{+1}-\theta_{-1})}_{\mathbf{w}}\tau(x)+\underbrace{\frac{a_0(\theta_{+1}).p}{a_0(\theta_{-1})(1-p)}}_{b}\right) \tag{3}$$

$$=sign(\langle\mathbf{w},\tau(x)\rangle+b)$$

This result is very interesting as it shows that, for a rather large class of generative models, the final classification function is a linear function in the model parameters $\theta=(\theta_{-1},\theta_{+1},p)$. Now consider the special case that the distribution $\mathbf{P}_{X|Y=y,Q=\theta}$ of objects $x\in\mathcal{X}$ given classes $y\in\{-1,+1\}$ is a multidimensional Gaussian in some feature space $K\subseteq\ell_2^n$ mapped into by some given feature map $\phi:\mathcal{X}\to K$,

$$\mathbf{f}_{X|Y=y,Q=\theta}(x)=(2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mu)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_y)\right) \tag{4}$$

Where the parameters $\boldsymbol{\theta}_y$ are the mean vector $\boldsymbol{\mu}_y\in\mathbb{R}^n$ and the covariance matrix $\boldsymbol{\Sigma}_y\in\mathbb{R}^{n\times n}$, respectively. Making the additional assumption that the covariance matrix $\boldsymbol{\Sigma}$ is the same for both models $\boldsymbol{\theta}_{+1}$, $\boldsymbol{\theta}_{-1}$ and $\mathbf{P}_{Y|Q=\theta}(+1)=\mathbf{P}_{Y|Q=\theta}(-1)$ we see that,

$$\boldsymbol{\theta}=\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu};-\frac{\Sigma^{-1}_{11}}{2};-\Sigma^{-1}_{12};....;\frac{\Sigma^{-1}_{22}}{2};-\Sigma^{-1}_{23};....;-\frac{\Sigma^{-1}_{nn}}{2}\right) \tag{5}$$

$$\boldsymbol{\tau}(\mathbf{x})=(\mathbf{x};x_1^2;x_1x_2;....;x_1x_n;x_2^2;x_2x_3;....;x_n^2)$$

$$\tau_0(x)=1$$

$$a_0(\boldsymbol{\theta})=(2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}\right) \quad (6)$$

according to equations (3, 5, and 6) then,

$$\boldsymbol{\tau}(x)=\mathbf{x},\quad \mathbf{w}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{+1}-\boldsymbol{\mu}_{-1}),\ b=\frac{1}{2}\left(\boldsymbol{\mu}'_{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{-1}-\boldsymbol{\mu}'_{+1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{+1}\right) \quad (7)$$

This result also follows from substituting (4) directly in to equation (2) (see Figure 1**: (left)** The black line represents the decision boundary. This must always be a linear function because both models use the same (estimated) covariance matrix $\hat{\Sigma}$ (ellipses).

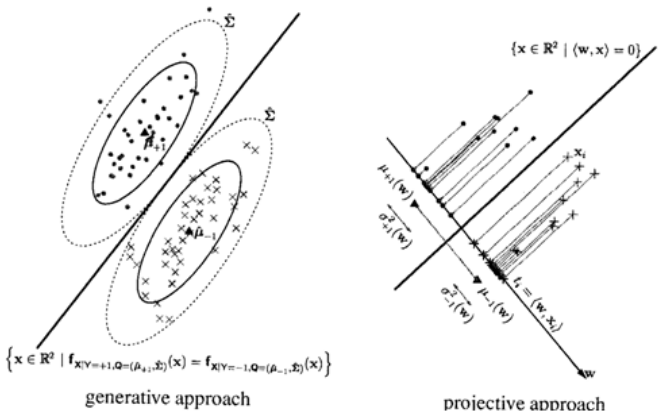

Figure 1: Fisher Discriminant

An appealing feature of this classifier is that it has a clear geometrical interpretation which was proposed for the first time by R. A. Fisher. Instead of working with n-dimensional vectors $\mathbf{x}$ we consider only their projection onto a hyperplane with normal $\mathbf{w}\in \mathrm{K}$. Let $\boldsymbol{\mu}_y(\mathbf{w})=\mathbf{E}_{\mathrm{X}|\mathrm{Y}=y}\left[\mathbf{w}'\phi(\mathrm{X})\right]$ be the expectation of the projections of mapped objects $\mathbf{x}$ from class y onto the linear Discriminant having normal $\mathbf{w}$ and $\sigma_y^2(\mathbf{w})=\mathbf{E}_{\mathrm{X}|\mathrm{Y}=y}\left[\left(\mathbf{w}'\phi(\mathrm{X})-\boldsymbol{\mu}_y(\mathbf{w})\right)^2\right]$ the variance of these projections. Then choose as the direction $\mathbf{w}\in \mathrm{K}$ of the linear Discriminant a direction along which the maximum of the relative distance between the $\boldsymbol{\mu}_y(\mathbf{w})$ is obtained, that is, the direction $\mathbf{w}_{FD}$ along which the maximum of

$$J(w)=\frac{\left(\mu_{+1}(\mathbf{w})-\mu_{-1}(\mathbf{w})\right)^2}{\sigma_{+1}^2(\mathbf{w})+\sigma_{-1}^2(\mathbf{w})} \quad (8)$$

is attained. Intuitively, the numerator measures the inter class distance of points from the two classes $\{+1,-1\}$ whereas the denominator measures the intra-class distance points in each of the two classes see also Figure (1) right, that a

geometrical interpretation of the Fisher Discriminant objective function (8), given a weight vector $\mathbf{w} \in K$, each mapped training object $\mathbf{x}$ is projected onto $\mathbf{w}$ by virtue of $t = \langle \mathbf{x}, \mathbf{w} \rangle$. The objective function measures the ratio of the inter-class distance $(\mu_{+1}(\mathbf{w}) - \mu_{-1}(\mathbf{w}))^2$ and the intra-class distance $\sigma_{+1}^2(\mathbf{w}) + \sigma_{-1}^2(\mathbf{w})$.Thus the function J is maximized if the inter-class distance is large and the intra-class distance is small. In general, the Fisher linear Discriminant $\mathbf{w}_{FD}$ suffer from the problem that its determination is a very difficult mathematical and algorithmical problem, However, in the particular case of $\mathbf{P}_{\mathsf{X}|\mathsf{Y}=y,\mathsf{Q}=\theta} = Normal(\mu_y, \Sigma)^2$, a closed form solution to this problem is obtained by noticing that $\mathsf{T} = \mathrm{w}'\phi(\mathsf{X})$ is also normally distributed with $\mathbf{P}_{\mathsf{T}|\mathsf{Y}=y,\mathsf{Q}=\theta} = Normal(\mathbf{w}'\mu_y, \mathbf{w}'\Sigma\mathbf{w})$. Thus, the objective function given in equation (8) can be written as

$$J(w) = \frac{(w'(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1}))^2}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} + \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}} = \frac{1}{2} \cdot \frac{\mathbf{w}'(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1})(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1})'\mathbf{w}}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}} \tag{9}$$

Which is known as the generalized Rayleigh quotient having the maximizer $\mathbf{w}_{FD}$,

$$\mathbf{w}_{FD} = \Sigma^{-1}(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1}) \tag{10}$$

This expression equals the weight vector $\mathbf{w}$ found by considering the optimal classification under the assumption of a multidimensional Gaussian measure for the class conditional distributions $\mathbf{P}_{\mathsf{X}|\mathsf{Y}=y}$

Unfortunately, as with the discriminative approach, we do not know the

parameters $\theta = (\boldsymbol{\mu}_{+1}, \boldsymbol{\mu}_{-1}, \boldsymbol{\Sigma}) \in \mathcal{Q}$ but have to "learn" them from the given training sample $z = (\mathrm{x}, \mathrm{y}) \in \mathcal{Z}^{\mathrm{m}}$. We shall employ the Bayesian idea of expressing our prior belief in certain parameters via some prior measure $\mathbf{P}_{\mathsf{Q}}$. After having seen the training sample z we update our prior belief $\mathbf{P}_{\mathsf{Q}}$, giving a posterior belief $\mathbf{P}_{\mathsf{Q}|\mathsf{Z}^{\mathrm{m}}=z}$. Since we need one particular parameter value we compute the MAP estimate $\hat{\boldsymbol{\theta}}$, that is, we choose the value of $\boldsymbol{\theta}$ which attains the maximum a-posterior belief $\mathbf{P}_{\mathsf{Q}|\mathsf{Z}^{\mathrm{m}}=z}$[3]. If we choose a (improper) uniform prior $\mathbf{P}_{\mathsf{Q}}$ then the parameter $\hat{\boldsymbol{\theta}}$ equals the parameter vector which maximize the likelihood and is therefore also known as the maximum likelihood estimator, these estimates are given by

$$\hat{\boldsymbol{\mu}}_y = \frac{1}{m_y} \sum_{(x_i, y) \in z} \mathbf{x}_i \quad , \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{2} \sum_{y \in \{+1,-1\}} \sum_{(x_i, y) \in z} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_y)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_y)' \tag{11}$$

$$= \frac{1}{m}\left(\mathbf{x}'\mathbf{x} - \sum_{y \in \{+1,-1\}} m_y \hat{\boldsymbol{\mu}}_y \hat{\boldsymbol{\mu}}'_y \right)$$

[2] Note that $\mu_y(\mathbf{w}) \in \mathbb{R}$ is a real number whereas $\mu_y(\mathbf{w}) \in \mathbb{R}$ is an a –dimension vector in feature space .

[3]For details see Linear Kernel Classifiers p.80.

Where is the data matrix obtained by applying $\phi : \mathcal{X} \to \mathbf{K}$ to each training object $x \in \mathbf{x}$ and m_y equals the number of training examples of class y. Substituting the estimates into the equations (7) results in the so-called Fisher Linear Discriminant $\mathbf{w}_{FD}$. The pseudo code of this algorithm is given in Appendix (A) **(6, 10)**

KERNEL FISHER DISCRIMINANT

In an attempt to "kernelize" the algorithm of Fisher Linear Discriminant its note that a crucial requirement is that $\hat{\mathbf{\Sigma}} \in \mathbb{R}^{n \times n}$ has full rank which is impossible if $\dim(\mathrm{K}) = n \succ\succ m$. Since the idea of using kernels reduces computational complexity in these cases we see that it is impossible to apply a kernel trick directly to this algorithm. Therefore, let us proceed along the following route: Given the data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ we project the m data vectors $\mathbf{x}_i \in \mathbb{R}^n$ into the m-dimensional space spanned by the mapped training objects $\mathbf{x} \longrightarrow \mathbf{X}\mathbf{x}$ and then estimate the mean vector and the covariance matrix in $\mathbb{R}^m$ using equation (11). The problem with this approach is that $\hat{\mathbf{\Sigma}}$ is at most of rank $m-2$ because it is un outer product matrix of two centered vectors. In order to remedy this situation we apply the technique of regularization to the resulting of $m \times m$ covariance matrix, i.e. we penalize the diagonal of this matrix by adding $\lambda \mathbf{I}$ to it where large value of λ corresponds to increasing penalization. As a consequence, the projection m-dimensional mean vector $\mathbf{k}_y \in \mathbb{R}^m$ and covariance matrix $\mathbf{S} \in \mathbb{R}^{m \times m}$ are given by

$$\mathbf{k}_y = \frac{1}{m_y} \sum_{(x_i, y) \in z} \mathbf{X}\mathbf{x}_i = \frac{1}{m_y} \mathbf{G}\left(\mathbf{l}_{y_1 = y}, \ldots . \mathbf{l}_{y_m = y}\right)'$$

$$\mathbf{S} = \frac{1}{m}\left(\mathbf{X}\mathbf{X}'\mathbf{X}\mathbf{X}' - \sum_{y \in \{+1, -1\}} m_y \mathbf{k}_y \mathbf{k}_y{}'\right) + \lambda \mathbf{I}$$

$$= \frac{1}{m}\left(\mathbf{G}\mathbf{G} - \sum_{y \in \{+1, -1\}} m_y \mathbf{k}_y \mathbf{k}_y{}'\right) + \lambda \mathbf{I}$$

Where the $m \times m$ matrix $\mathbf{G}$ with $\mathbf{G}_{ij} = \langle x_i, x_j \rangle = k(x_i, x_j)$ is the Gram matrix. Using $\mathbf{k}_y$ and $\mathbf{S}$ in place of $\boldsymbol{\mu}_y$ and $\mathbf{\Sigma}$ in the equations (7) results so-called kernel Fisher Discriminant. We note that the m-dimensional vector computed corresponds to the linear expansion coefficients $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^m$ of a weight vector $\mathbf{w}_{KFD}$ in feature space because the classification of a novel test object $x \in \mathcal{X}$ by the Kernel Fisher Discriminant is carried out on the projected data point $\mathbf{X}\mathbf{x}$,i.e.

$$h(x) = sign\left(\langle \hat{\boldsymbol{\alpha}}, \mathbf{X}\mathbf{x} \rangle + \hat{b}\right) = sign\left(\sum_{i=1}^{m} \hat{\alpha}_i k(x_i, x) + \hat{b}\right)$$

$$\hat{\boldsymbol{\alpha}} = \mathbf{S}^{-1}\left(\mathbf{k}_{+1} - \mathbf{k}_{-1}\right), \quad \hat{b} = \frac{1}{2}\left(\mathbf{k}'_{-1}\mathbf{S}^{-1}\mathbf{k}'_{-1} - \mathbf{k}'_{+1}\mathbf{S}^{-1}\mathbf{k}'_{+1}\right) \tag{12}$$

It is worth mentioning that we would have obtained the same solution by exploiting the fact that the objective function (8) depends only on inner products between mapped training objects $\mathbf{x}_i$ and the unknown weight vector $\mathbf{w}$. By virtue of **Representer Theorem**[4] the solution can be written as $\mathbf{w}_{FD} = \sum_{i=1}^{m} \hat{\boldsymbol{\alpha}}_i \mathbf{x}_i$ which inserted into (8), yields a function in $\boldsymbol{\alpha}$ whose maximizer is given by equation (8). the pseudocode of this algorithm is given in Appendix (B).[6,10]

RAYLEIGH COEFFICIENT

To find the optimal linear Discriminant we need to maximize a Rayleigh coefficient (cf. Equation (9)). Fisher's Discriminant can also be interpreted as a feature extraction technique is defined by the separability criterion (8). From this point of view, we can think of the Rayleigh coefficient as a general tool to fined features which (i) cover much of what is considered to be interesting (e.g. variance in PCA), (ii) and at the same time avoid what is considered disturbing (e.g. within class variance in Fisher's Discriminant). The ratio in (9) is maximized when one covers as much as possible of the desired information while avoiding the undesired. We have already shown in Fishers Discriminant that this problem can be solved via a generalized eigenproblem. By using the same technique, one can also compute second, third, etc., generalized eigenvectors from the generalized eigenproblem, for example in PCA where we are usually looking for more than just one feature.[10]

REGULARIZATION

The optimizing Rayleigh coefficient for Fisher's Discriminant in a feature space poses some problems. For example if the matrix Σ is not strictly positive and numerical problems can cause the matrix Σ not even to be positive semi-definite. Furthermore, we know that for successful learning it is

[4]for details see learning Kernel Classifiers p.48

absolutely mandatory to control the size and complexity of the function class we choose our estimates from. This issue was not particularly problematic for linear Discriminant since they already present a rather simple hypothesis class. Now, using the kernel trick, we can represent an extremely rich class of possible non-linear solutions, we can always achieve a solution with zero within class variance (i.e. $\mathbf{w}'\Sigma\mathbf{w}$). Such a solution will, except for pathological cases, be over fitting.

To impose a certain regularity, the simplest possible solution to add a multiple of the identity matrix to Σ, i.e. replace Σ by Σ_λ where

$$\Sigma_\lambda = \Sigma + \lambda I \qquad (\lambda \geq 0)$$

This Can Be Viewed in Different Ways

- If λ is **sufficiently large** this makes the problem feasible and numerically more stable as Σ_λ becomes positive definite.
- **Increasing** λ decreases the variance inherent to the estimate Σ; for $\lambda \to \infty$ the estimate become less and less sensitive to the covariance structure. In fact, for $\lambda = \infty$ the solution will lie in the direction of $\mathbf{m}_2 - \mathbf{m}_1$. The estimate of this "means" however, converges very fast and is very stable [(2)].
- For a **suitable choice of** λ**,** this can be seen as decreasing the bias in sample based estimation of eigenvalue [(4)]. The point here is, that the empirical eigenvalue of a covariance matrix is not an unbiased estimator of the corresponding eigenvalue of the true covariance matrix, i.e. as we see more and more examples, the largest eigenvalue does not converge to the largest eigenvalue of the true covariance matrix. One can show that the largest eigenvalues are over-estimated and that the smallest eigenvalues are under-estimated

However, the sum of all eigenvalues (i.e. the trace) does converge since the estimation of the covariance matrix itself (when done properly) is unbiased.

Another possible regularization strategy would be to add a multiple of the kernel matrix K *to* S, i.e. replace S with

$$\Sigma_\lambda = \Sigma + \lambda K \qquad (\lambda \geq 0)$$

The regularization value compute by using Cross-Validation, the pseudocode of this algorithms given in Appendix (C)

PRACTICAL PART

Introduction

In this paper By taking a small sample size 20 observation we want clear how these observations classified by testing this classified using statistic (Rayleigh Coefficient), the concepts is maximizing the distance between group means with minimizing the distance within groups to obtain the optimal solution by using one of non-linear Fisher Discriminant its Kernel Fisher Discriminant In primal and dual variable with two levels (± 1).

From Appendix (A), Fisher Discriminant in primal variable by taking $\lambda = \sigma^2$ =0.57373 computed by Generalize (leave one out) cross-validation, see Appendix (C) have a vector of coefficients i.e.

$$\mathbf{w} = \begin{bmatrix} -0.000365865255644 \\ 0.701870728355830 \\ 0.000002878439743 \\ 0.127346290424504 \\ 0.000174843224781 \\ -0.006478024238546 \end{bmatrix}$$

$$J(\mathbf{w}) = \frac{15.306248465897635}{1.956160043675995}$$
$$= 7.824640174703855$$

b = 2.513798662130466

From the value of statistic Rayleigh coefficient with primal variable clear that the distance between groups greater than the distance within groups means the separate of between-class scatter matrix is maximized and the within-class scatter matrix is minimized that is the required and the solution is feasible

From Appendix (B), Fisher Discriminant in dual variable we obtain on vector of coefficients by taking $\lambda = \sigma^2$ =0.57373 computed by Generalize (leave one out) cross-validation, see Appendix (C) have a vector of coefficients i.e.

$$\boldsymbol{\alpha} = \begin{bmatrix} 0.295508396750165 \\ -0.066421330540834 \\ -0.185588761869894 \\ 0.207219565728792 \\ -0.172647343149947 \\ -0.091287131039280 \\ 0.153758948367567 \\ -0.126371566473608 \\ -0.145210858850987 \\ -0.203239090402064 \\ -0.071985084958214 \\ 0.289147653036707 \\ -0.082999970338278 \\ 0.309758211838471 \\ -0.066195126644743 \\ 0.214846717137107 \\ -0.042010264371129 \\ -0.173937016302716 \\ -0.131837467187324 \\ -0.068577206984628 \end{bmatrix}$$

$$J(\boldsymbol{\alpha}) = \frac{22.788451144064311}{.165849182106410}$$
$$= 1.374046640124104e + 002$$

b = -0.366755860887264

From the value of statistic Rayleigh coefficient with dual variable clear that the distance between groups greater

than the distance within groups means the separate of between-class scatter matrix is maximized and the within-class scatter matrix is minimized that is the required and the solution is feasible

CONCLUSIONS

By taking a small sample 20 case about HIV disease with three factors (Age, Gender, number of Lymphocyte cell) with two levels, the value of statistic Rayleigh coefficient with both primal and dual variables clear that the distance between groups greater than the distance within groups means the separate of between-class scatter matrix is maximized and the within-class scatter matrix is minimized, since both primal and dual solution are feasible then there exist an optimal (finite) solution, means the patients are classified in correct.

REFERENCES

1. Aronszajn. N. “Theory of Reproducing Kernels”. Transactions of American Mathematical Society, 68: 337-404, 1950.
2. Bousquet, O. and A. Elisseeff. “Stability and generalization”. Journal of Machine
3. Learning Research, 2:499–526, March 2002.
4. Fisher, A.R. “The use of multiple measurements in taxonomic problems.” Annals of Eugenics, 7:179-188, 1936.
5. Friedman, J.H. “Regularized discriminant analysis”. Journal of the American Statistical
6. Association, 84(405):165–175, 1989.
7. Fukunaga K.,” Introduction to Statistical Pattern Recognition ”, Academic press, INC,2nd ed, 1990
8. Herbrich, H. “Learning Kernel Classifiers.” Theory and Application. MIT Press, Cambridge 2002.
9. Kim et al., “Optimal Kernel Selection in Kernel Fisher Discriminant Analysis.” Department of Electrical Engineering, Stanford University, Stanford, CA 94304 USA.
10. Mika et al., “Fisher Discriminant Analysis with Kernels.” In Neural networks for signal processing IX, E.J. Larson and S.Douglas, eds., IEEE, 1999, pp. 41-48.
11. Mika et al., “A mathematical Programming Approach to the Kernel Fisher Algorithm.” In Advances in Neural Information Processing Systems, 13, pp. 591-579, MIT Press. 2001Heme
12. Mika S. “Kernel Fisher Discriminants.” PhD thesis, University of Technology, Berlin, 2002. [24]
13. Scholkopf and A.J. Smola. “learning with Kernels.” MIT Press, Cambridge, MA, 2002
14. Shawe-Taylor, J.,& Cristianini, N. “Kernel Methods for Pattern Analysis.” Cambridge: Cambridge University Press. 2004
15. Trevor Hastie et al.,” The Elements of Statistical learning. Data Mining, Inference, and prediction.” Second Edition. © Springer Science + Business Media, LLC (2009).

16. Yang et al., "KPCA plus LDA: A complete Kernel Fisher Discriminant Framework for Feature Extraction and Recognition." IEEE Transactions on Pattern Recognition and Machine Intelligence, 27, 230-244, 1989.

APPENDICES

Appendix (A)

Pseudo code for (Fisher Discriminant Analysis in primal variable)

Require: A feature mapping $\phi : \mathcal{X} \to \mathrm{K} \subseteq \ell_2^{\mathrm{n}}$

Require: A training sample $z = ((x_1, y_1), ..., (x_m, y_m))$

Determine the number m_{+1} and m_{-1} of samples of class $+1 \; and \; -1$

$$\hat{\mu}_{+1} = \frac{1}{m_{+1}} \sum_{(x_i,+1)\in z} \boldsymbol{\varphi}(x_i) \quad ; \quad \hat{\mu}_{-1} = \frac{1}{m_{-1}} \sum_{(x_i,-1)\in z} \boldsymbol{\varphi}(x_i)$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{m}\left(\sum_{i=1}^{m} \boldsymbol{\varphi}(x_i)\boldsymbol{\varphi}(x_i)' - m_{+1}\hat{\boldsymbol{\mu}}_{+1}\hat{\boldsymbol{\mu}}_{+1} - m_{-1}\hat{\boldsymbol{\mu}}_{-1}\hat{\boldsymbol{\mu}}_{-1} \right) + \lambda \mathbf{I}_m$$

$$\mathbf{w} = \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_{+1} - \hat{\boldsymbol{\mu}}_{-1})$$

$$J(w) = \frac{(w'(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1}))^2}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} + \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}} = \frac{1}{2} \cdot \frac{\mathbf{w}'(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1})(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1})'\mathbf{w}}{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}}$$

$$b = \frac{1}{2}\left(\hat{\boldsymbol{\mu}}'_{-1}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_{-1} - \hat{\boldsymbol{\mu}}'_{-1}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_{-1} \right)$$

Appendix (B)

Pseudo code for (Fisher Discriminant Analysis in dual variable)

Require: A training sample $z = ((x_1, y_1), ..., (x_m, y_m))$

Require: A kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and regularization parameter $\lambda \in \mathbb{R}^{+}$

Determine the number m_{+1} and m_{-1} of samples of class $+1 \; and \; -1$

$$G = \left(k(x_i, x_j)\right)_{i,j=1}^{m,m} \in \mathbb{R}^{\mathrm{m\times m}}$$

$$k_{+1} = \frac{1}{m_{+1}} G\left(\mathbf{I}_{\mathrm{y_1}=+1}, ..., \mathbf{I}_{\mathrm{y_m}=+1}\right)' \; ; \; k_{-1} = \frac{1}{m_{-1}} G\left(\mathbf{I}_{\mathrm{y_1}=-1}, ..., \mathbf{I}_{\mathrm{y_m}=-1}\right)'$$

$$\mathbf{S} = \frac{1}{m}\left(\mathbf{GG} - m_{+1}\mathbf{k}_{+1}\mathbf{k}'_{+1} - m_{-1}\mathbf{k}_{-1}\mathbf{k}'_{-1}\right) + \lambda \mathbf{I}_m$$

$$\boldsymbol{\alpha} = \mathbf{S}^{-1}\left(\mathbf{k}_{+1} - \mathbf{k}_{-1}\right)$$

$$J(w) = \frac{1}{2} \cdot \frac{\mathbf{w}'\left(\mathbf{k}_{+1} - \mathbf{k}_{-1}\right)\left(\mathbf{k}_{+1} - \mathbf{k}_{-1}\right)' \mathbf{w}}{\mathbf{w}'\mathbf{S}\mathbf{w}}$$

$$b = \frac{1}{2}\left(\mathbf{k}_{-1}\mathbf{S}^{-1}\mathbf{k}_{-1} - \mathbf{k}_{+1}\mathbf{S}^{-1}\mathbf{k}_{+1}\right) + \ln\left(\frac{m_{+1}}{m_{-1}}\right)$$

return the vector $\boldsymbol{\alpha}$ of expansion coefficients and offset $b \in \mathbb{R}$

Appendix C

Algorithm for cross-validation **(13)**

$$S = \mathbf{x}\left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'$$

$$\hat{f}(x_i) = \hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$$

$$GCV\left(\hat{f}\right) = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \hat{f}(x_i)}{1 - trace(S)/N}\right]^2$$