12. SVM & Flexible Discriminants

오영민

Introduction

- **Review for LDA: Linear Discriminant Analysis**
- **CCA: Canonical Correlation Analysis**
- **LDA & CCA**
- **FDA: Flexible Discriminant Analysis**

Review for LDA

- For training set $\{x_i, g_i\}_1^N$ where $x_i \in \mathbb{R}^p$, $g_i \in \mathcal{G} = \{1,...,K\}$,

$$\hat{G}(x) = \max_k \hat{P}r(G = k \,|\, X = x) = \max_k \frac{\hat{P}r(X = x \,|\, G = k)Pr(G = k)}{\hat{P}r(X = x)} = \max_k \frac{\hat{f}_k(x)\hat{\pi}_k}{\sum \hat{\pi}_k f_k(x)}$$

where $\hat{\pi}_k = N_k/N$, $X|_{G=k} \sim \mathcal{N}(\hat{\mu}_k, \hat{\Sigma}_W)$, $\hat{\mu}_k = \sum_{g_i=k} x_i/N_k$, $\hat{\Sigma}_W = \sum_k \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T/(N-K)$, $\forall k$

- Assume $\hat{\pi}_k$ are same for all k, then $\hat{G}(x) = \min_k (x - \hat{\mu}_k)^T \hat{\Sigma}_W^{-1}(x - \hat{\mu}_k)$. i.e. $\hat{G}(x)$ is classified to the class with centroid closest to x, where distance is measured in the **Mahalanobis metric** using the pooled within group covariance matrix $\hat{\Sigma}_W$.
- The decision boundaries created by LDA: $\log \frac{f_k(x)}{f_l(x)} = x^T \hat{\Sigma}_W^{-1}(\hat{\mu}_k - \hat{\mu}_l) + C = 0$ which is linear in x.

Review for LDA

- **(Sphering: Mahalanobis to Euclidean metric).** Since Σ_W is symmetric, SVD of Σ_W is $UDU^T = \|\sqrt{D}U\|^2$. Now for input x, let $x* = UD^{-1/2}x$. Then $Var(x*) = D^{-1/2}U^T\Sigma_W UD^{-1/2} = I_p$ i.e. The decision rule is $\arg min_k \|x* - \mu_k^*\|^2$.
  Denote $\Sigma_W^{-1/2} = D^{-1/2}U^T$. Then $\arg min_k \|x* - \mu_k^*\|^2 = \arg min_k \|\Sigma_W^{-1/2^T}(x - \mu_k)\|^2$

- **LDA provides natural low-dimensional views of the data:** Since $\mu_k = \mu_K + (\mu_k - \mu_K)$ for $k = 1,...,K-1$, the K-centroids in $\mathbb{R}^p$ lies in affine subspace of dimension at most $K-1$, denote H_{K-1}.

- **(PCA, Optimal scoring)** Moreover, we can get $L < K-1$-dimensional subspace $H_L \subset H_{K-1}$ **optimal** for LDA. In other words, the projected centroids were spread out as much as possible.
  - Compute the covariance matrix of $\{\mu_1^*, \ldots, \mu_K^*\}$, Σ_B^* and also compute its eigen vector, eigen value matrix $V*, D_B$, respectively. ($D_B = diag(d_1, \ldots, d_K)$). Then d_l be the l-th largest eigen value and corresponding to the eigen vector v_l^*.
  - $v_l = \Sigma_W^{-1/2} v_l^*$ is called **the l-th canonical or discriminant vector.** Let $U = (v_1 .,,, v_L)$. Then $H_L = \{U^T x : x \in \mathbb{R}^p\}$.

Review for LDA

- **Summary:**

1. Gaussian classification with common covariances leads to linear decision boundaries

2. Since only the relative distances to the centroids count, one can confine the data to the subspace spanned by the centroids in the sphered space H_K.

3. H_K can be further decomposed into successively optimal subspaces $H_L \subset H_K$ in term of centroid separation. The reduced subspaces have been motivated as a data reduction (for viewing) tool and also be used for classification.

CCA: Canonical Correlation Analysis

- **We can recast LDA as a regression problem.**

- Let $Y \in \mathbb{R}^{N\times K}$ be one-hot encoded target vector in training set and suppose that $\theta : \mathcal{G} \to \mathbb{R}$ is a function that assigns scores to the classes s.t. $\theta(\cdot)$'s are optimally predicted by linear regression on $X \in \mathbb{R}^{N\times p}$. e.g. a linear map $\eta(x) = x^T\beta$.

- In general, we can find L-sets of independent scorings $\{\theta_1, \ldots \theta_L\}$ and L-corresponding linear maps $\eta(x) = x^T\beta_l$, $\forall l$.

- $\Theta = (\theta_1, \ldots, \theta_L) \in \mathbb{R}^{K\times L}$ where $\theta_l = (\theta_l(1), \ldots, \theta_l(K))^T \in \mathbb{R}^K$, $l = 1,\ldots,L$ and $B = (\beta_1, \ldots, \beta_L)$, $\beta_l \in \mathbb{R}^p$, $\forall l$ are the parameters for CCA. Our goal is to find the optimal (Θ, B) pairs that minimize:

$$ASR = \sum_{l=1}^{L}\sum_{i=1}^{N}(\theta_l(g_i) - x_i^T\beta_l)^2/N = tr(\|Y\Theta - XB)\|^2)/N$$

- Note. Let $\Theta^* = Y\Theta$, then $\{\Theta^*\}_{il} = \theta_l(g_i)$ and $\{\|Y\Theta - XB)\|^2\}_{lk} = \sum_{i=1}^{N}(\theta_l(g_i) - x_i^T\beta_l)(\theta_k(g_i) - x_i^T\beta_k)$.

- $\theta_l^{*T}\theta_l^* = 1$, $\forall l$ **and** $\theta_l^{*T}\theta_k^* = 0$, $\forall l \neq k$ **to prevent trivial 0 solutions.**

CCA: Canonical Correlation Analysis

- For fixed Θ, $\hat{\beta}_l = (X^TX)^{-1}X^T\theta_l^*$, $\forall l$ i.e. $\hat{B} = (X^TX)^{-1}X^T\Theta^*$.

- **Theorem.** The sequence of canonical vectors v_l is **identical** to the sequence of β_l up to a constant.

- Then for $\hat{\beta}$, the optimization problem is minimizes $ASR(\Theta) = tr(\|\Theta^* - X\hat{B})\|^2)/N = tr(\Theta^{*^T}(I - P_X)\Theta^*)/N$ subject to $\Theta^{*^T}\Theta^* = I_L$ (or $\theta_l^{*^T}\theta_l^* = 1$, $\forall l$ and $\theta_l^{*^T}\theta_k^* = 0$, $\forall l \neq k$) i.e. $L + L(L-1)/2$ - constraints.

- Then the constraint term in Lagrange multiplier is $\lambda_1(\theta_1^{*^T}\theta_1^* - 1) + \ldots + \lambda_L(\theta_L^{*^T}\theta_L^* - 1) + \lambda_1'(\theta_1^{*^T}\theta_2^* - 0) + \lambda_1'(\theta_2^{*^T}\theta_1^* - 0) + \ldots + \lambda_{L(L-1)/2}'(\theta_{L-1}^{*^T}\theta_L^* - 0) = tr(\Lambda(\Theta^{*^T}\Theta^* - I_L))$ where Λ is **symmetric** with positive elements.

- Since Λ is symmetric, SVD of Λ is $V\Sigma V^T$ where $V^TV = I$ i.e. $\mathscr{L} = tr(\Theta^{*^T}(I - P_X)\Theta^*)/N - tr(\Lambda(\Theta^{*^T}\Theta^* - I_L))$ = $tr(M^T(I - P_X)M) - tr(\Sigma(M^TM - I_L))$ where $M = \Theta^*V$. **As both Λ and Σ are dummy variables and can have any name, we can initially assume that Λ is diagonal.**

- **Consider the optimization problem:** $\min_{\Theta} \left[tr(\Theta^TY^T(I - P_X)Y\Theta)/N - tr(\Lambda(\Theta^TY^TY\Theta - I_L))\right]$

CCA: Canonical Correlation Analysis

- $\frac{\partial}{\partial\Theta}\left[tr(\Theta^T Y^T(I - P_X)Y\Theta)/N - tr(\Lambda(\Theta^T Y^T Y\Theta - I_L))\right] = 2Y^T(I - P_X)Y\Theta/N - 2Y^T Y\Theta\Lambda = 0$

- $Y^T Y\Theta(I/N - \Lambda) = Y^T P_X Y\Theta/N$. Since Y is one-hot encoded, $Y^T Y = diag(N_1, \ldots, N_K)$ i.e. $(Y^T Y)^{-1} = diag(1/N_1, \ldots, 1/N_K)$. Denote $\Lambda' = (I_L/N - \Lambda) = diag(1/N - \lambda_1, \ldots, 1/N - \lambda_L)$. Then, $\Theta\Lambda' = (Y^T Y)^{-1} Y^T P_X Y/N\Theta$. Therefore, Θ is eigen vector matrix of $(Y^T Y)^{-1} Y^T P_X Y/N$ corresponding to eigen value matrix Λ'.

- Specifically, $U^T x = (x^T v_1, \ldots x^T v_L)^T = (d_1 x^T \beta_1, \ldots, d_L x^T \beta_L)^T = DB^T x$ where $d_l = 1/\alpha_l^2(1 - \alpha_l^2)$ and α_l is the l-th largest eigen values computed in Λ'.

CCA: Canonical Correlation Analysis

- **LDA by optimal scoring** (Thus, $L = K$ or $\Theta \in \mathbb{R}^{K \times K}$ case):

    1. **Initialize.** Form one-hot encoded vector $Y \in \mathbb{R}^{N \times K}$ and set $\Theta_0 = I_K$ i.e.$\Theta_0^* = Y$.
    2. **Multivariate regression.** Regress $\Theta^* = Y\Theta$ on X; Set $\hat{Y} = P_X Y = XB$ where $B \in \mathbb{R}^{p \times K}$ is the coefficient matrix.
    3. **Optimal scores.** Obtain the eigen vector matrix Θ of $(Y^T Y)^{-1} Y^T P_X Y / N = (Y^T Y)^{-1} Y^T \hat{Y} / N = (Y^T Y)^{-1} Y^T XB / N$
    4. **Update.** Update the coefficients $B \leftarrow B\Theta$.

- In above procedure, we compute $Y^T P_X Y$ without explicitly computing P_X itself.

FDA: Flexible Discriminant Analysis

- Then we can generalize $\eta(x) = x^T B$ to $h(x)^T B$ where $h(x) = (h_1(x), \ldots, h_M(x)) \in \mathbb{R}^m$ e.g. MARS, BRUTO, PPR, NNs…

- When the non-parametric regression procedure can be represented as a **linear operator**; $\hat{Y} = S_\lambda Y$, then the procedures of FDA is same as LDA by optimal scoring with one change: Replace P_X with S_λ

- **Additive splines** have this property, if the smoothing promoters λ are fixed: MARS, BRUTO.

- After Initialize - Multivariate regression - Optimal scores - Update step, we can get the optimal fit $\eta(\cdot)$ and fitted class centroids $\bar{\eta}^k = \sum_{g_i = k} \eta(x_i)/N_k, \forall k$.

- For input x, the decision rule is: $\delta(x, k) = ||D(\eta(x) - \bar{\eta}^k)||^2$. Note. $\eta(x)$ has at most $(K-1)$-elements.

PDA: Penalized Discriminant Analysis

- For some classes of problems, involving the basis expansion, is not needed; we already have far **too many (correlated) predictors**. e.g. spoken speech data, image data,…

- Positively correlated predictors lead to noisy, negatively correlated coefficients estimates, and this noise results in un-wanted sampling variance. A reasonable strategy is to regularize the coefficients to be smooth over the spatial domain.

- e.g. Consider MARS procedure ($f(x) = \alpha + f_1(x_1) + \ldots, f_p(x_p)$ using spline), the optimization problem is:

$$\frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N}[\theta_l(g_i) - \sum_{j=1}^{p} f_{lj}(x_{ij})^2] + \sum_{l=1}^{L}\sum_{j=1}^{p}\lambda_j \int f''_{lj}(t)^2 dt$$

where λ_j is roughness penalty for the j-term.
(**trade off between fit and smoothness.**)

Note. λ_j are same for L-models i.e. Non-parametric regression must be able to handle a multiple response variables when selecting λ.

- Then we know that solution is a finite dimensional form: $f_{lj}(x_j) = h_{jl}(x_j)^T\beta_{jl}$ with $\beta_{lj} \in \mathbb{R}^N$.

- And the optimization problem is : $\|\Theta^* - HB\|^2 + B^T\Omega B$ where $\Omega_\lambda = diag(\lambda_1\Omega_1, \ldots, \lambda_p\Omega_p) \in \mathbb{R}^{Np\times Np}$, $\Omega_j \in \mathbb{R}^{N\times N}\ \forall j$,
$B = (\beta_1, \ldots, \beta_L) \in \mathbb{R}^{Np\times K}$, $\beta_l = (\beta_{l1}^T, \ldots, \beta_{lp}^T)^T \in \mathbb{R}^{Np}\ \forall l$, $H = (h(x_1), \ldots, h(x_N))^T \in \mathbb{R}^{N\times Np}$, $h(x_i) = (h_1(x_i), \ldots, h_{Np}(x_i))^T\ \forall i$.

- Then the regression operator has the form: $S_\lambda = H(H^TH + \Omega_\lambda)^{-1}H^T$ and Θ minimizes $tr(\Theta^TY^T(I - S_\lambda)Y\Theta)/N$.

PDA: Penalized Discriminant Analysis

- This optimization problem corresponding to a form of PDA: Penalized Discriminant Analysis.

- Let Σ_B be the between-group covariance matrix for $h(x)$ and let $\Sigma_W + \Omega$ be the penalized within-group covariance matrix. Then, we define: **A PDA finds a matrix U to maximize $tr(U^T \Sigma_B U)$ subject to $U^T(\Sigma_W + \Omega)U = I$.**

- Using Lagrange multiplier, $\Sigma_B U = (\Sigma_W + \Omega)U\Lambda$, $U\Lambda^{-1} = (\Sigma_B + \epsilon I)^{-1}(\Sigma_W + \Omega)U$ where $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$. Thus, U is approximately the eigen vector matrix for $(\Sigma_B + \epsilon I)^{-1}(\Sigma_W + \Omega)$.

- And the penalized Mahalanobis distances from class centroids in the augmented space of $h(x)$ is given by:

$$\delta(x, k) = (h(x) - \bar{h}^k)^T(\Sigma_W + \Omega)^{-1}(h(x) - \bar{h}^k) = ||D(\eta(x) - \bar{\eta}^k)||^2$$

- Loosely speaking, the penalized Mahalanobis distance tends to give less weight to “rough” coordinates, and more weight to “smooth” ones.

PDA: Penalized Discriminant Analysis

- **Model selection: Choosing λ via cross validation**

$$GCV(c, \lambda) = \frac{ASR(\lambda)}{[1 - \{1 + c \cdot df(\lambda)\}/N]^2}$$

- $df(\lambda)$ is the effective degrees of freedom in the model. For MARS, $df(\lambda)$ is the number of independent basis functions, whreas for a BRUTO it measures the amount of smoothing. In both cases, $df(\lambda) = tr(S_\lambda) - 1$.

- c is called the cost per degree of freedom. Based on the work of Friedman(1991) and Owen (1991), it seems that reasonable values are 2 for additive models(BRUTO, degree-1 MARS), and 3 for higher-degree MARS.

MDA: Mixture Discriminant Analysis

- LDA can be viewed as a **prototype** classifier: Each class is represented by its class centroid, and we classify to the closest using an appropriate metric.; In many cases, a single prototype is insufficient to represent inhomogeneous classes.

- e.g. GMM for the k-th class has density: $Pr(X|G=k) = \sum_{r=1}^{R_k} \pi_{kr}\phi(X;\mu_{kr},\Sigma)$ subject to $\pi_{k1} + \ldots + \pi_{kR_k} = 1,\ \forall k$. Then the decision rule is given by:

$$\arg max_k Pr(G=K|X=x) = \arg max_k \sum_{r=1}^{R_k} \pi_{kr}\phi(x;\mu_{kr},\Sigma)\Pi_k \text{ where } \Pi_k = Pr(G=k).$$

- Given R_ks, we estimate the set of parameters $\theta = \{\pi_{kr}, \mu_{kr}, \Pi_k, \Sigma\}$; $(2*(R_1+\ldots+R_K)+K+p(p+1)/2)$-parameters. Often Π_k are known or proportion in trining data. Thus, set $\theta = \{\pi_{kr}, \mu_{kr}, \Sigma\}$.

$$\arg max_\theta\ l(\theta) \text{ where } l(\theta) = \underline{\sum_k \sum_{g_i=k} \log[\sum_{r=1}^{R_k} \pi_{kr}\phi(x_i;\mu_{kr},\Sigma)\Pi_k]}$$

Sum within the log form i.e. hard to optimize directly
→ Use EM algorithms.

MDA: Mixture Discriminant Analysis

- **(E-step)** Given the current parameters, compute the responsibility of subclass c_{kr} within class k:

For each class k, $$W(c_{kr} \mid x_i, g_i) = \frac{\pi_{kr}\phi(x_i; \mu_{kr}, \Sigma)}{\sum_{l=1}^{R_k} \pi_{kl}\phi(x_i; \mu_{kl}, \Sigma)}$$ where $r = 1, \ldots, R_k$.

- **(M-step)** Compute the weighted MLEs for the parameters of each of the component Gaussians within each of the classes, using the weights from the E-step.

For each class k, compute $$\hat{\pi}_{kr} \propto \sum_{g_i = k} W(c_{kr} \mid x_i, g_i)$$ subject to $$\sum_r \hat{\pi}_{kr} = 1,$$

$$\hat{\mu}_{kr} = \frac{\sum_{g_i=k} W(c_{kr} \mid x_i, g_i) x_i}{\sum_{g_i=k} W(c_{kr} \mid x_i, g_i)}$$ where $r = 1, \ldots, R_k$,

$$\hat{\Sigma} = \frac{1}{N-K} \sum_k \sum_{g_i=k} \sum_r W(c_{kr} \mid x_i, g_i)(x_i - \hat{\mu}_{kr})(x_i - \hat{\mu}_{kr})^T$$

MDA: Mixture Discriminant Analysis

- **Model selection: Choosing R_k, and initial values of $W(c_{kr}|x_i, g_i)$, $\{\mu_{kr}\}$, Σ via k-means clustering:**

  For each class k, we choose a fixed number of clusters R_k and then use k-means clustering to estimate $\{\mu_{kr}\}$. Then for all observations in class k, $W(c_{kr}|x_i, g_i)$ is set to 1 if μ_{kr} is closest centroid to x_i and to 0 o.t.

- **MDA by optimal scoring:** Optimal scoring procedure is carries over to the M-step of the MDA. Instead of using one-hot encoded response $Y \in \mathbb{R}^{N \times K}$, we use blurred response $Z \in \mathbb{R}^{N \times R}$ where $R = R_1 + \ldots + R_K$, whose row consist of the current subclass probabilities for each observation.

  1. **Initialization** via k-means clustering.

  2. **Multivariate regression** Z on X : $\hat{Z}$ be the fitted values and H be the vector of fitted regression functions.

  3. **Optimal scoring:** Obtain the eigen vector matrix Θ of $(Z^T Z)^{-1} Z^T S_\lambda Z$.

  4. **Update** $B = B\Theta$ and $W(c_{kr}|x_i, g_i)$, π_{kr}'s in M-step.