Linear Classifiers

A linear classifier for the two-class 0-1 coded problem is given by

$$x \mapsto x^T\beta + \beta_0$$

with the classifier at x_0

$$f_{\beta_0,\beta}(x) = \begin{cases} 1 & \text{if } x^T\beta + \beta_0 \geq \frac{1}{2} \\ 0 & \text{if } x^T\beta + \beta_0 < \frac{1}{2} \end{cases}$$

With $(x_1, y_1), \ldots, (x_N, y_N)$ a data set we can minimize the average empirical 0-1-loss

$$(\beta_0, \beta) \mapsto \sum_{i=1}^{N} 1(y_i \neq f_{\beta_0,\beta}(x_i)))$$

Not easy, discontinuous, solution not unique. View $x^T\beta + \beta_0$ as a local model of $P_x(1)$ and consider

$$\operatorname*{argmin}_{\beta_0,\beta} \sum_{i=1}^{N} (y_i - x_i^T\beta - \beta_0)^2.$$

One-dimensional Normal Variables

Let X be real valued and $X|Y = k$ be $N(\mu_k, \sigma^2)$ for $k = 0, 1$. If $\Pr(Y = k) = \pi_k$ the Bayes classifier is

$$f(x) = \begin{cases} 0 & \text{if } \pi_0 g_0(x) \geq \pi_1 g_1(x) \\ 1 & \text{if } \pi_0 g_0(x) < \pi_1 g_1(x) \end{cases}$$

Or

$$f(x) = \begin{cases} 0 & \text{if } \log(g_0(x)/g_1(x)) \geq \log(\pi_1/\pi_0) \\ 1 & \text{if } \log(g_0(x)/g_1(x)) < \log(\pi_1/\pi_0) \end{cases}$$

Or

$$f(x) = \begin{cases} 0 & \text{if } 2x(\mu_0 - \mu_1) \geq 2\sigma^2 \log(\pi_1/\pi_0) - \mu_1^2 + \mu_0^2 \\ 1 & \text{if } 2x(\mu_0 - \mu_1) < 2\sigma^2 \log(\pi_1/\pi_0) - \mu_1^2 + \mu_0^2 \end{cases}$$

Linear Discriminant Analysis

Let Y take values in $\{1, \ldots, K\}$ with

$$\Pr(Y = k) = \pi_k$$

with $\pi_1 + \ldots + \pi_K = 1$, and let the conditional distribution of $X|Y = k$ be $N(\mu_k, \Sigma)$ on $\mathbb{R}^p$ with Σ regular. That is, the density for $X|Y = k$ is

$$g_k(x) = \frac{1}{\sqrt{2\pi \det(\Sigma)}^p} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}.$$

The conditional probability of $Y = k|X = x$ is

$$\Pr(Y = k|X = x) = \frac{\pi_k g_k(x)}{\pi_1 g_1(x) + \ldots + \pi_k g_1(x)}$$

The Bayes Classifier

$$
\begin{aligned}
\log \frac{\Pr(Y = k|X = x)}{\Pr(Y = l|X = x)} &= \log \frac{\pi_k}{\pi_l} + \log \frac{g_k(x)}{g_l(x)} \\
&= \log \frac{\pi_k}{\pi_l} + \frac{1}{2}(x - \mu_l)^T \Sigma^{-1}(x - \mu_l) - \frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k) \\
&= \log \frac{\pi_k}{\pi_l} + \frac{1}{2}\mu_l^T \Sigma^{-1}\mu_l - \frac{1}{2}\mu_k^T \Sigma^{-1}\mu_k + x^T \Sigma^{-1}(\mu_k - \mu_l)
\end{aligned}
$$

The boundary – the x's where $\Pr(Y = k|X = x) = \Pr(Y = l|X = x)$ – is a hyperplane. We call this a linear classifier as we can determine the classification by the computation of the finite number of linear functions $x^T \Sigma^{-1}(\mu_k - \mu_l)$, $k, l = 1, \ldots, K$.

Linear Discriminant Functions

Introducing

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

we see that

$$\log \frac{\Pr(Y = k | X = x)}{\Pr(Y = l | X = x)} = \delta_k(x) - \delta_l(x)$$

The decision boundaries are the solutions to the linear equations

$$\delta_k(x) = \delta_l(x)$$

and the Bayes classifier is

$$f(x) = \text{argmax}_k \delta_k(x).$$

Figure 4.5 – Linear Discrimination

Estimation

We use the the plug-in principle for estimation. That is, maximum likelihood estimation of all the parameters in the full model for (X, Y)

$$
\begin{aligned}
\hat{\pi}_k &= \frac{N_k}{N}, \quad N_k = \sum_{i=1}^{N} 1(y_i = k) \\
\hat{\mu}_k &= \frac{1}{N_k} \sum_{i:y_i=k} x_i \\
\hat{\Sigma} &= \frac{1}{N-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T
\end{aligned}
$$

– with the usual centralized estimate of the covariance matrix.

Lecture exercise

Solve lecture exercise 2.

Parameter Functions

Fixing the last group K as a reference group we have for $k = 1, \ldots, K-1$ that

$$\log \frac{\Pr(Y = k | X = x)}{\Pr(Y = K | X = x)} = \underbrace{\log \frac{\pi_k}{\pi_K} + \frac{1}{2}\mu_K^T \Sigma^{-1} \mu_K - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k}_{\beta_{k0}} + x^T \underbrace{\Sigma^{-1}(\mu_k - \mu_K)}_{\beta_k}$$

Thus

$$\Pr(Y = k | X = x) = \frac{\exp(\beta_{k0} + x^T \beta_k)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + x^T \beta_l)}$$

for $k = 1, \ldots, K-1$. The conditional distribution depends upon $\pi_1, \ldots, \pi_{K-1}, \mu_1, \ldots, \mu_k, \Sigma$ through the parameter function

$$(\pi_1, \ldots, \pi_{K-1}, \mu_1, \ldots, \mu_K, \Sigma) \mapsto (\beta_{10}, \ldots, \beta_{(K-1)0}, \beta_1, \ldots, \beta_{K-1}).$$

Estimation Methodology – a digression

Non-model based (the direct) approach:

- Local methods aiming directly for (non-parametric) estimates of e.g. $E(Y \mid X = x)$ or $P(Y = k \mid X = x)$.
  Example: Nearest neighbors.
- Empirical risk minimization: Take $\mathcal{F}$ to be a set of predictor functions and take
  $$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum L(y_i, f(x_i)).$$
  Example: Least squares fit of linear regression and classification models.

Estimation Methodology – a digression

Introduce a parametrized statistical model $(P_\theta)_{\theta \in \Theta}$ of the generating probability distribution.

Model based approach

- The plug-in principle: If $\hat{\theta}$ is an estimator of θ and f_θ is the optimal predictor under P_θ take $f_{\hat{\theta}}$.
  Example: LDA.
- The conditional plug-in principle: Assume that the conditional distribution, $P_{x,\tau(\theta)}$, of Y given $X = x$ depends upon θ through a parameter function $\tau : \Theta \to \Theta_2$. Then $f_\theta = f_{\tau(\theta)}$ and if $\hat{\tau}$ is an estimator of τ we take $f_{\hat{\tau}}$.
  Examples: Model based linear regression and logistic regression.

Quadratic Discriminant Analysis

What if $\Sigma_1 \neq \Sigma_2$ ($K = 2$)?

$$\log \frac{\Pr(Y = k | X = x)}{\Pr(Y = l | X = x)} \quad = \quad \bar{\delta}_k(x) - \bar{\delta}_l(x)$$

where

$$\bar{\delta}_k(x) = -\frac{1}{2} \log \det \Sigma_k - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) + \log(\pi_k).$$

is a quadratic function. The decision boundaries are the solutions to the quadratic equations $\bar{\delta}_k(x) = \bar{\delta}_l(x)$ and the Bayes classifier is

$$f(x) = \text{argmax}_k \bar{\delta}_k(x).$$

Figure 4.6 – Quadratic Discrimination

To get quadratic boundaries one can either do QDA (right) or one can transform the bivariate variable $X = (X_1, X_2)^T$ to the five dimensional variable $X' = (X_1, X_2, X_1^2, X_1 X_2, X_2^2)$ and do LDA in $\mathbb{R}^5$ (left). The linear boundary in $\mathbb{R}^5$ shows up as a quadratic boundary in $\mathbb{R}^2$.

Figure 4.4 – Dimension Reduction

Linear discriminant analysis provides a direct dimension reduction to the K-dimensional space. The above figure shows a further reduction to a 2D projection chosen to maximize the spread of the group means.

Figure 4.9 – Discrimination and Dimension Reduction

How to project to maximize the spread of group means? The usual inner product in Euclidean space is not optimal – we should use the inner product given by Σ^{-1}

Change of Basis Point of View

If $\Sigma = cVD^2V^T$ with D a diagonal matrix with strictly positive entries and $c > 0$ we let $\tilde{x} = D^{-1}V^T x$ and $\tilde{\mu}_k = D^{-1}V^T\mu_k$. This is a change of basis given by the matrix $D^{-1}V^T$. With R a constant not depending on k we have

$$\begin{aligned} \log \Pr(Y = k|X = x) &= \log \pi_k - \frac{1}{2c}(x - \mu_k)^T VD^{-2}V^T(x - \mu_k) + R \\ &= \log \pi_k - \frac{||\tilde{x} - \tilde{\mu}_k||^2}{2c} + R. \end{aligned}$$

Hence

$$\text{argmax}_k \Pr(Y = k|X = x) = \underset{k}{\text{argmin}} \left(||\tilde{x} - \tilde{\mu}_k||^2 - 2c\log \pi_k\right).$$

LDA as Dimension Reduction Technique

With W_0 a "sphering" matrix fulfilling that

$$\hat{\Sigma} = cW_0^T W_0$$

the empirical covariance matrix of the "sphered" data $\tilde{x}_k = W_0^{-1} x_k$ is cI.

- Take M^* to be the $K \times p$ matrix of class means of the "sphered" data $\tilde{x}_k$.
- Take $B^* = V^*(D^*)^2(V^*)^T$ to be the covariance matrix of M^*.

Then the columns in V^*, ordered decreasingly according to the diagonal entries in D^*, form an orthonomal basis (canonical variates) in the "sphered" coordinates.

Figure 4.8 – Dimension Reduction