Review: The Bayes Classifier

For the discrete loss matrix \( L = [L(k, \ell)] \),

\[
EPE = E[L(Y, f(X))] = E_x \left\{ \sum_{k=1}^{K} L(k, f(X))P(Y = k|X) \right\} ;
\]

we can minimize pointwise to obtain

\[
f(x) = \arg\min_{\ell} \sum_{k=1}^{K} L(k, \ell)P(Y = k|X = x) .
\]

When \( L(k, \ell) = 1_{k \neq \ell} \), zero-one loss, we obtain

\[
f(x) = \arg\min_{\ell} \sum_{k \neq \ell} P(Y = k|X = x)
\]

\[
= \arg\min_{\ell} \{1 - p(Y = \ell|X = x)\}
\]

\[
= \arg\max_{\ell} P(Y = \ell|X = x)
\]

This the the posterior mode, or Bayes classifier. Its error rate is the Bayes rate.
Linear and Quadratic Discriminant Analysis and Friends

- All these methods produce classification functions $\delta_k(x)$ with which we classify $x$ into class $k = \arg\max_k \delta_k(x)$.

- Generalization of our earlier linear regression method: Linear regression of an indicator matrix; $\delta_k(x)$ will be linear.

- Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are variations on a Gaussian mixture model that uses training data to define the clusters (hence no E-M needed).
  - LDA: Common variance/covariance matrix $\Sigma$ within each cluster; $\delta_k(x)$ is linear in $x$.
  - QDA: Different $\Sigma_k$ in each cluster $k$; $\delta_k(x)$ is quadratic in $x$.

- Logistic regression leads to similar classification rules ($\delta_k(x)$ is again linear) but treats the relationship between data and model differently.

---

Linear regression of an indicator matrix

- Let $Y_{n \times K} = [y_{ij}]$ be a class-indicator matrix; that is, there are $n$ rows for the $n$ training observations, and each row has $K - 1$ 0’s and one 1 indicating the class of the observation.

- Let $X_{n \times p}$ be the matrix of “features” (including a column of 1’s).

- Then $\hat{B} = (X^T X)^{-1} X^T Y$ is a matrix of linear regression coefficients, one column for each column of $Y$.

- For a new row-vector observation $x$, the fitted values $\delta(x) = [\delta_1(x), \ldots, \delta_K(x)] = x \hat{B}$ provide a classification rule:

  $$\text{Class}(x) = \arg\max_k \delta_k(x)$$
Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA)

Here we start more explicitly from the Bayes classifier, which chooses according to

$$\hat{\text{Class}}(x) = \arg \max_k P[G_i = k | x_i = x]$$

where \(G_i\) is the “true” classification of observation \(i\). If \(\pi_k\) are the prior probabilities of the \(K\) classes (\(P[G_i = k] = \pi_k\)), then clearly

$$P[G_i = k | x_i = x] = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

and, as we said before, our main task is to model the within-class densities \(f_k(x)\).

LDA, QDA and related methods begin with the Gaussian model

$$f_k(x) = \frac{1}{(2\pi)^{p/2}\sigma_k^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$$

(and so these are the “classification” analogues to Gaussian-mixture-model cluster analysis).

Some interpretations:

- It is easy to verify that \(\delta(x_i) = x_i\hat{B}\) for the regression coefficients \(\hat{B}\) minimize the sum of squares

$$\min_{\hat{B}} \sum_{i=1}^{n} ||y_i - x_i\hat{B}||^2;$$

and for a new observation \(x\), classifying according to

$$\hat{\text{Class}}(x) = \arg \min_k ||\delta(x) - t_k||^2,$$

where \(t_k\) is the \(k^{th}\) row of \(I_K \times K\) (identity matrix), is the same as classifying according to \(\text{Class}(x) = \arg \max_k \delta_k(x)\).

- Can also view the linear regression functions \(\delta_k(x)\) here as crude approximations to \(P(\text{obs } i \text{ belongs to class } k | x_i)\) in the Bayes classifier.

- In either case, increasing the dimension of \(x\) should improve the classification; e.g.:
  - Polynomial functions of \(X\);
  - Projecting \(X\) onto an appropriate orthogonal set of basis functions;
  - etc.
**Linear discriminant analysis (LDA)**

LDA arises in the special case that $\Sigma_k \equiv \Sigma$, $\forall k$. In comparing two classes $k$ and $\ell$ it is enough to look at the log-ratio of posterior probabilities

$$\log \frac{P[G = k|X = x]}{P[G = \ell|X = x]} = \log \frac{f_k(x)}{f_\ell(x)} + \log \frac{\pi_k}{\pi_\ell}$$

$$= \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2} \left( (\mu_k + \mu_\ell)^T \Sigma^{-1} (\mu_k - \mu_\ell) + x^T \Sigma^{-1} (\mu_k - \mu_\ell) \right)$$

after some algebra and cancellation due to $\Sigma_k \equiv \Sigma$. This in turn leads to the **linear discriminant functions**

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

In practice we use unbiased estimates of the cluster parameters:

- $\hat{\pi}_k = N_k/N$, the fraction of training examples in class $k$;
- $\hat{\mu}_k = \frac{1}{N_k} \sum_{G_i = k} x_i$;
- $\Sigma = \frac{1}{N-K} \sum_{k=1}^K \sum_{G_i=k} (x_i - \hat{\mu}_k) (x_i - \mu_k)^T$.

**Quadratic discriminant analysis (QDA)**

With Quadratic discriminant analysis we allow $\Sigma_k$ to vary between clusters. There are then fewer cancellations in the algebra above, and we are lead to **quadratic discriminant functions**

$$-\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + \log \pi_k$$

Some observations/comparisons:

- LDA and linear regression of the indicator matrix $Y$ correspond exactly in the case of two classes.
- Both LDA and linear regression of the indicator matrix, with polynomial functions of $X$, perform similarly to QDA with simply linear $X$ (this might be expected!).
- When the assumption of Gaussian mixtures is correct, the above unbiased estimates of parameters lead to optimal (empirical Bayes) classification rules.
- When the assumption of Gaussian is not correct, one can generally improve the classifications in by adding biases $b_k$ to the discrimination function, i.e.,
  $$\delta'_k(x) = \delta_k(x) + b_k$$
  to optimize classification loss in the training data.
Since LDA also assumes a linear form for the log ratios,

\[
\log \frac{P[G = k|X = x]}{P[G = K|X = x]} = \beta_{k0} + \beta_k^T x
\]

which is better/preferable? Let us write,

\[
P(X, G = k) = Pr(X)P[G = k|X]
\]

- Logistic regression estimates \(P[G = k|X]\) conditional on \(X\), ignoring \(Pr(X)\);
- LDA estimates the full model \(P(X, G = k) = P(G = k)P[X|G = k]\), using a specific model for \(P[X|G = k]\)

If the Gaussian model for \(P[X|G = k]\) is correct, LDA will do better. In other settings, logistic regression is thought to provide a more robust classifier. However, in most circumstances, LDA and logistic regression provide very similar answers.
A nearly linearly-separable case...

linear regression of indicators

Ida, orig features

Ida, quadratic features

qda, orig features

Ida, degree 4 polynomial features

lda, degree 4 polynomial features

err rate = 0.045

Ida, degree 2 polynomial features

lda, degree 2 polynomial features

err rate = 0.03

Ida, degree 1 polynomial features

lda, degree 1 polynomial features

err rate = 0.045

Ida, degree 3 polynomial features

lda, degree 3 polynomial features

err rate = 0.04

Ida, degree 4 polynomial features

lda, degree 4 polynomial features

err rate = 0.03
A quadratically-separable case...

linear regression of indicators

Ida, orig features

er rate = 0.419354837096777

lda, quadratic features

qda, orig features

er rate = 0.209677419354839

lda, degree 4 polynomial features

Ida, degree 2 polynomial features

er rate = 0.145161290322581

er rate = 0.209677419354839

lda, degree 3 polynomial features

Ida, degree 4 polynomial features

er rate = 0.145161290322581

er rate = 0.06451612903258

er rate = 0