Classification Problems:

- Classification Problems
  - Training data: \( \{(x_1, g_1), (x_2, g_2), \ldots, (x_n, g_n)\} \), the responses are categorical.
  - The goal is to form the \( G(X) \) to predict \( G \) based on \( X \). \( G \) takes values in \( \mathcal{G} = \{G_1, G_2, \ldots, G_k\} \)
  - Example: Email spam classification, handwritten digits recognition

- \( G(X) \) divides the input space into a collection of regions, each labeled by one class.

Linear Methods for Classification

- Linear methods: decision boundaries are linear

- For a two class problem, the decision boundary is

  \[
  \left\{ x : \alpha_0 + \sum_{j=1}^{p} \alpha_j x_j = 0 \right\}
  \]

  The two regions separated by this hyperplane: \( \left\{ x : \alpha_0 + \sum_{j=1}^{p} \alpha_j x_j > 0 \right\} \) and \( \left\{ x : \alpha_0 + \sum_{j=1}^{p} \alpha_j x_j < 0 \right\} \)

- For more than two classes, the decision boundary between any pair of classes is a hyperplane and therefore the whole input space is piecewise hyperplanar.

Bayes classifier:

- Let \( \hat{G} \) be the estimated classifier for \( G \)

- Define the loss function \( L(k, l) \) as the price paid for classifying an observation belonging to class \( k \) to class \( l \)

  Let \( L(k, l) = 0 \) if \( k = l \); otherwise 0 - this is called zero-one loss function.

- The expected prediction error is

  \[
  EPE = E \left[ L \left( G, \hat{G}(X) \right) \right]
  \]

  the expectation is taken with respect to the joint distribution \( \Pr(G, X) \).
• The Bayes classifier:

\[ \hat{G}(X) = G_k \text{ if } \Pr(G_k|X = x) = \max_{g \in \mathcal{G}} \Pr(g|X = x) \]

This says we classify to the most probable class, using conditional distribution \( \Pr(G|X) \).

• The error rate of the Bayes classifier is called the Bayes rate.

**Linear Regression of an Indicator Matrix**

• For a \( K \) class classification problem, there will be \( K \) class indicators \( Y_k, k = 1, 2, \ldots, K \).

\[ Y_k = 1 \text{ if } G = k \text{ else } 0 \]

• Fit a linear regression model for each \( Y_k \) using \( X \):

\[ \hat{\beta}^k = (X^TX)^{-1}X^TY_k \text{ and } \hat{y}_k = X\hat{\beta}^k \]

• Let \( \hat{\mathbf{B}} = (\hat{\beta}^1, \ldots, \hat{\beta}^k) = (X^TX)^{-1}X^T\mathbf{Y} \), for a new observation

\[ \hat{f}(x) = \left[ (1, x)^T \hat{\mathbf{B}} \right]^T \]

\[ = \begin{pmatrix} \hat{f}_1(x) \\ \hat{f}_2(x) \\ \vdots \\ \hat{f}_k(x) \end{pmatrix} \]

• Identify the largest component and classify accordingly:

\[ \hat{G}(x) = \arg \max_{k \in \mathcal{G}} \hat{f}_k(x) \]

**The Rationale**

• View the regression as an estimate of conditional expectation: For the random variable \( Y_k \),

\[ E(Y_k|X = x) = P(G = k|X = x) \]

• The rational is

– Approximate \( P(G = k|X = x) \) by a linear function of \( x \) using linear regression.

– Apply the Bayes rule to the approximated probability.
• The question is: how well a linear regression model approximates the conditional probability?

• We have $\sum_{k \in G} \hat{f}_k(x) = 1$ for any $x$, as long as there is an intercept in the model, but the $\hat{f}_k(x)$ can be negative or greater than 1 due to the rigid nature of linear regression. This problem itself does not guarantee that this approach will not work.

• Masking Problem: There is a serious problem with the regression approach when the number of classes $K \geq 3$, especially prevalent when $K$ is large. Because of the rigid nature of the regression model, classes can be masked by others.
Linear Discriminant Analysis

- Decision Theory for classification implies we need to know the class posteriors \( P(G|X) \) for optimal classification.

- Let \( \pi_k \) the prior probability of class \( k \) and \( f_k (x) \) the class conditional density of \( X \) in class \( G = k \), then the posterior probability can be computed,

\[
Pr (G = k|X = x) = \frac{f_k (x) \pi_k}{\sum_{l=1}^{k} f_l (x) \pi_l}
\]

- By Bayes Rule

\[
\hat{G} (x) = \arg\max_k Pr (G = k|X = x)
= \arg\max_k f_k (x) \pi_k
\]

- Suppose we model each class density as multivariate Gaussian

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

- Linear discriminant analysis (LDA): \( \Sigma_k = \Sigma, \forall k \). We see that the linear discriminant functions

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

are an equivalent description of the decision rule, with

\[
G (x) = \arg\max_k \delta_k (x)
\]

- The decision boundary between class \( k \) and \( l \) is

\[
\{ x : \delta_k (x) = \delta_l (x) \}
\]

Or equivalent the following holds

\[
\log \frac{\pi_k}{\pi_l} - \frac{1}{2} (\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + x^T \Sigma^{-1} (\mu_k - \mu_l) = 0
\]

The boundary is linear in \( x \).
Estimate Gaussian Distributions

- In practice, we need to estimate the Gaussian distribution
- \( \hat{\pi}_k = N_k / N \), where \( N_k \) is the number of class-\( k \) samples.
- \( \hat{\mu}_k = \frac{\sum_{g_i=k} x_i}{N_k} \)
- \( \hat{\Sigma} = \frac{\sum_{k=1}^K \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T}{(N - k)} \)

The Connection between LDA and Linear Regression (of the class indicator matrix)

- Two class problem: LDA direction is the same as Least Squares direction when \( N_1 = N_2 \).
- For more than two classes, LDA is not the same as linear regression of the class indicator matrix, and it avoids the the masking problem.

Quadratic Discriminant Analysis (QDA)

- If the \( \Sigma_k \) are not assumed to be equal, then the convenient cancellations do not occur. We then get quadratic discriminant functions,

\[
\delta_k(x) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma^{-1}_k (x - \mu_k) + \log \pi_k
\]

- The decision boundary between each pair of classes \( k \) and \( l \) is described by a quadratic equation \( \{x : \delta_k(x) = \delta_l(x)\} \)

QDA and LDA

- The number of parameters need to be compute: LDA \( \sim (K - 1) \times (p + 1) \); QDA \( \sim (K - 1) \times \{p(p + 3) / 2 + 1\} \)
- Both LDA and QDA perform well on an amazingly large and diverse set of classification tasks, probably because the data can only support simple decision boundaries such as linear or quadratic, and the estimates provided via the Gaussian models are stable.
- For LDA, we can put up with the bias of a linear decision boundary because it can be estimated with much lower variance than more exotic alternatives.

Regularized Discriminant Analysis

- A compromise between LDA and QDA
• Regularized covariance matrices:

\[ \hat{\Sigma}_k(\alpha) = \alpha \hat{\Sigma}_k + (1 - \alpha) \hat{\Sigma}, \alpha \in [0, 1] \]

• "Shrink" the separate covariance of QDA toward the pool covariance as in LDA.
• The quadratic discriminant function \( \delta_k(x) \) is defined using the regularized covariance matrices \( \hat{\Sigma}_k(\alpha) \).
• \( \alpha \) controls the complexity of the model and can be chosen by cross-validation. (Fig 4.7)

Feb 11, 2008

Computation for QDA (and LDA)

• To compute the discriminant function:

\[ \delta_k(x) = -\frac{1}{2} \log |\hat{\Sigma}_k| - \frac{1}{2} (x - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (x - \hat{\mu}_k) + \log \hat{\pi}_k \]

• Diagonalize \( \hat{\Sigma}_k \): \( \hat{\Sigma}_k = V_k D_k^2 V_k^T \). \( V_k \) is \( p \times p \) orthonormal, and \( D_k \) a diagonal matrix of non-negative eigenvalues \( d_{kl} \).

• Then

\[ (x - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (x - \hat{\mu}_k) = [D_k^{-1} V_k^T (x - \hat{\mu}_k)]^T [D_k^{-1} V_k^T (x - \hat{\mu}_k)] \]

and

\[ \log |\hat{\Sigma}_k| = 2 \sum l \log d_{kl} \]

• LDA

  – Sphere the data with respect the common covariance estimate \( \hat{\Sigma} \). \( X^* \leftarrow D^{-1} V^T X \)

  – Classify to the closest class centroid with \( \log \pi_k \) adjusted.

Reduced-Rank Linear Discriminant Analysis

• Part of LDA’s popularity is due to an additional restriction that allows us to view informative low-dimensional projections of the data.
• The subspace spanned by the \( K \) centroids is at most of rank \( K - 1 \), denoted by \( H_{K-1} \).
• \( X^* \) can be projected onto \( H_{K-1} \) without affecting the classification result. Thus there is a fundamental dimension reduction in LDA, namely that we only need to consider the data in a subspace of dimension at most \( K - 1 \).
• If $K = 3$, this could allow us to view the data in the two-dimensional plot, color-coding the classes.

• What if $K > 3$? We might want to find a subspace $H_{K-1}$, $L < K - 1$ optimal for LDA in some senses.

**Optimization Criterion**

• Fisher’s optimization criterion: the projected centroids were spread out as much as possible in terms of variance.

• With Gaussian distribution assumption, this amounts to finding the principal component subspaces of the centroids themselves.

• Fisher method (without Gaussian distribution assumption): Find the linear combination $Z = a^T X$ such that between class variance is maximized relative to the within-class variance.

• Within-class variance $W = \hat{\Sigma}$;

  Between-class variance $B$:

  $$\hat{\mu} = \sum_{k=1}^{K} \hat{\pi}_k \hat{\mu}_k$$

  $$B = \sum_{k=1}^{K} \hat{\pi}_k (\hat{\mu}_k - \hat{\mu})(\hat{\mu}_k - \hat{\mu})^T$$

• For $Z$, the within class variance is $a^T W a$ and between class variance is $a^T B a$.

• Fisher’s optimization (Fig 4.9) becomes

  $$\max_a \frac{a^T B a}{a^T W a}$$

  or equivalently

  $$\max_a a^T B a \text{ subject to } a^T W a = 1$$

• $a_1$ can be obtained by solving an eigen-decomposition problem: $a_1$ is given by the largest eigenvalue of $W^{-1} B$. Similarly one can find the next direction $a_2$, $a_2^T W a_1 = 0$, such that $a_2^T B a_2 / a_2^T W a_2$ is minimized. The $a_l$ are referred to as discriminant coordinates. (Fig 4.4, 4.8)
Summary

- Gaussian classification with common covariances leads to linear decision boundaries. Classification can be achieved by sphering the data with respect to $W \ (\hat{\Sigma})$, and classifying to the closest centroid (adjusted for $\log \pi_k$) in the sphered space.

- 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.

- This subspace can be further decomposed into successively optimal subspaces in terms of centroid separation. This decomposition is identical to the decomposition due to Fisher.

- The reduced subspace can be used for classification by limiting the distance-to-centroid calculations to the chosen subspace. One can show that this is a Guassian classification rule with the additional restriction that the centroids of the Gaussians lie in a $L$-dimensional subspace of $\mathbb{R}^p$ (Fig 4.11)
Logistic Regression

Motivation:

- Model the posterior distribution probabilities of the $K$ classes via linear functions in $x$
- Ensure that the probabilities sum to one and remain in $[0, 1]$.

Think about how to achieve this goal in a 2-class case...

- The probabilities $p_i = \Pr (G = 1|X = x_i)$ relate to the linear functions in $x$ through a link function, the logit (log-odds) function
  \[
  \log \frac{p_i}{1 - p_i} = \beta_0 + \beta^T x_i
  \]
- The 0-1 response $y_i$ is generated from the Bernoulli($p_i$).

The general $K$-class case: the model is specified in the terms of $K - 1$ logit transformations.

The log-likelihood for two-class case

\[
\begin{align*}
L (\theta) &= \prod_{i=1}^{N} p_{g_i} (x_i; \theta) \\
\end{align*}
\]

\[
\begin{align*}
l (\theta) &= \sum_{i=1}^{N} \log p_{g_i} (x_i; \theta) \\
&= \sum_{i=1}^{N} \{ y_i \log p (x_i; \beta) + (1 - y_i) \log (1 - p (x_i; \beta)) \} \\
&= \sum_{i=1}^{N} \left\{ y_i \log \left( \frac{p (x_i; \beta)}{1 - p (x_i; \beta)} \right) + \log \left( 1 - \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} \right) \right\} \\
&= \sum_{i=1}^{N} \left\{ y_i \beta^T x_i - \log \left( 1 + e^{\beta^T x_i} \right) \right\}
\end{align*}
\]
• To maximize the log-likelihood, we set its derivatives to zero.

\[
\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i (y_i - p(x_i; \beta)) = 0
\]

\[
= \sum_{i=1}^{N} x_i \left( y_i - \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} \right)
\]

These are \( p + 1 \) nonlinear equations in \( \beta \).

• To solve there equations, we use the Newton-Raphson algorithm, which requires the second-derivative or Hessian matrix

\[
\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{N} x_i x_i^T p(x_i; \beta) (1 - p(x_i; \beta))
\]

• Starting from \( \beta^{old} \), a single Newton-Raphson update is

\[
\beta^{new} = \beta^{old} - \left( \frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial l(\beta)}{\partial \beta}
\]

• Let \( X \) the \( N \times (p + 1) \) matrix of \( x_i \), \( p \) the vector of \( p(x_i; \beta) \) and \( W \) a \( N \times N \) diagonal matrix of weights with \( i \)th diagonal element \( p(x_i; \beta) (1 - p(x_i; \beta)) \), then

\[
\frac{\partial l(\beta)}{\partial \beta} = X^T (y - p)
\]

\[
\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = -X^T WX
\]

The Newton-Raphson step is then

\[
\beta^{new} = \beta^{old} + (X^T WX)^{-1} X^T (y - p)
\]

\[
= (X^T WX)^{-1} X^T W (X\beta^{old} + W^{-1} (y - p))
\]

\[
= (X^T WX)^{-1} X^T W z
\]

In the second and third line we have re-expressed the New-Raphson step as a weighted least squares step, with the response

\[
z = X\beta^{old} + W^{-1} (y - p)
\]

\[
\beta^{new} = \arg \min_{\beta} (z - X\beta)^T W (z - X\beta)
\]

• These equations get solved repeatedly, since at each iteration \( p \) changes, and hence so does \( W \) and \( z \). One usually starts with \( \beta = 0 \). Typically the algorithm does converge, since the log-likelihood is concave, but overshooting can occur, which can be avoided by size halving.
Logistic Regression or LDA?

- LDA: log-posterior odds between class k and K are linear functions of $x$.

$$\log \frac{\text{Pr}(G = k | X = x)}{\text{Pr}(G = K | X = x)} = \log \frac{\pi_k}{\pi_l} - \frac{1}{2} (\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + x^T \Sigma^{-1} (\mu_k - \mu_l)$$

$$= \alpha_{k0} + \alpha_k^T x.$$

This linearity is a consequence of the Gaussian assumption for the class density, as well as the assumption of a common covariance matrix.

- Logistic Regression: construct linear logits

$$\log \frac{\text{Pr}(G = k | X = x)}{\text{Pr}(G = K | X = x)} = \beta_{k0} + \beta_k^T x$$

- The same form; Difference lies in the way the linear coefficients are estimated.

- Logistic regression model is more general, in that it makes less assumptions, while LDA relies on Gaussian and common covariance assumption.

- As a result, LDA:
  - More stable: by relying on the additional model assumptions, we have more information about the parameters, and hence estimate them more efficiently (low variance).
  - Less robust to outliers: Observations that are far from the decision boundary play a role in estimating common covariance matrix. They are downweighted by logistic regression.

- In practice, LDA and logistic regression gives very similar result, even when LDA is used inappropriately, such as with qualitative predictors.