PROBABILISTIC BASED LINEAR MODELS

Supplement to
Bayesian Learning

Shai Fine
Linear Classification Methods

• Classes of linear methods for classification
  • Model the boundaries between classes
    • Rosenblatt’s Perceptron learning algorithm
    • Widrow-Hoff Least Mean Square (LMS) algorithm
    • MSE and Closed-form solution
  • Model a discriminant function for each class as linear
    • Linear discriminant analysis (LDA)
    • Logistic regression
Model Discriminant Function For Each Class

• Define a discriminant function $\delta_k(x): X \rightarrow \Omega$ for each class $k \in \Omega = \{\omega_1, \omega_2, ..., \omega_K\}$
  • The nature of $\delta_k(x)$ is based on the model

• Classification is via “the winner takes it all”
  $$g^*(x) = \arg \max_{\omega_k \in \Omega} \delta_k(x)$$

• The decision boundary between two classes
  $$\{x | \delta_k(x) = \delta_{k'}(x)\}$$
  • In case of linear discriminant functions this is a hyper-plane
Probabilistic-based Methods

- We will describe two linear classification methods which are based on
  - $\Pr(X = x|\omega_k)$ the conditional distribution of $x$ given the class $\omega_k$ – *likelihood*
  - $\Pr(\omega_k|X = x)$ the *posterior* probability of the class $\omega_k$ given $x$

- We will require that a monotone transformation of the discriminant function $\delta_k(x)$, will be linear

- Specifically, we will utilize the *logit* transformation, and require that
  \[
  \log \left[ \frac{p}{1 - p} \right] = \beta_0 + \beta^T x 
  \]

- Decision boundaries are the set of points with *log-odds*=0, and these are hyper-planes with \( \{x|\beta_0 + \beta^T x = 0\} \)

- Two popular methods that use log-odds
  - Linear Discriminant Analysis (LDA), Logistic Regression
Linear Projection

- A linear function (here in 2D)
  \[ f(x; w) = w_0 + x^T w \]
- Projects each point \( x = [x_1 \ x_2]^T \) to a line parallel to \( w \)

- We can study how well the projected points are separated across the classes, as a function of \( w \)
Projection and Classification

• By varying $\omega$ we get different levels of separation between the projected points.

• We would like to find a direction $\omega$ that maximizes the separation.
PCA vs. LDA

- **Principle Component Analysis (PCA)**
  - Find a low dimension useful to represent the data *irrespectively of class-membership*
  - The selected axes might not provide a good discriminative power
    - Class that are well separated in a high dimension, might produce mixtures on the projected axes

- **Linear Discriminative Analysis (LDA)**
  - Perform dimensionality reduction while preserving class discriminatory information
  - Looking for directions along which the classes are best separated
    - Utilizing class-membership information while moving and rotating the projected axes
Linear Discriminant Analysis (LDA)

**The Two-Class Case**

- The following are equivalent perspectives of the LDA transformation
  - **Maximize Rayleigh quotient**
    - Maximizes the between-class scatter while minimizing the within-class scatter
      \[
      \max \frac{w^T S_B w}{w^T S_W w}
      \]
    - $S_B$ – The scatter between-classes
    - $S_W$ – The scatter within-classes
  - **Maximize Fisher criterion**
    - Maximizes the distance between the cluster means, while scaling it by a measure of the within class scatter
      \[
      \max \frac{\left(\bar{m}_1 - \bar{m}_2\right)^2}{\hat{s}_1^2 + \hat{s}_2^2}
      \]
    - $\bar{m}_i, \hat{s}_i^2$ – The mean and scatter of class $i$ (resp.)
  - **Minimize Classification Error**
    - If the class conditional densities are multivarite with an equal covariance then LDA is equivalent to a classification by linear least squares
      \[
      w = \Sigma^{-1}(\mu_1 - \mu_2)
      \]
Scaled Distance Between Two Classes

- In the two classes setting, the projection is actually to a line
  \[ y = w^T x \]
- The distances between the projected means of the class
  \[ (\tilde{m}_1 - \tilde{m}_2)^2 = (w^T (m_1 - m_2))^2 \]
- Where
  - d-dim sample mean: \( m_k = 1/n_k \sum_{x \in x_k} x \)
  - 1-dim mean of the projected samples: \( \tilde{m}_k = 1/n_k \sum_{y \in y_k} y = 1/n_k \sum_{x \in x_k} w^T x \)
- The scatter for the projected samples
  \[ \tilde{s}_k^2 = \sum_{y \in y_k} (y - \tilde{m}_k)^2 \]
- Fisher criterion maximize the distance between the projected means scaled by the total within class scatter of the projected samples
  \[ \max \frac{\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \]
- Why scaling?
Original and Projected Space

- **Between-class scatter**

\[
(m_1 - m_2)^2 = (w^T (m_1 - m_2))^2 = w^T (m_1 - m_2)(m_1 - m_2)^T w = w^T S_B w
\]

- Where \(S_B = (m_1 - m_2)(m_1 - m_2)^T\) is the between-class covariance matrix (rank 1)

- **Within-class scatter**

\[
\begin{align*}
\hat{s}_1^2 + \hat{s}_2^2 &= \sum_{y \in y_1} (y - \bar{m}_1)^2 + \sum_{y \in y_2} (y - \bar{m}_2)^2 \\
&= \sum_{x \in x_1} w^T (x - m_1)(x - m_1)^T w + \sum_{x \in x_2} w^T (x - m_2)(x - m_2)^T w \\
&= w^T (S_1 + S_2) w \\
&= w^T S_W w
\end{align*}
\]

- Where \(S_i = \sum_{x \in x_i} (x - m_i)(x - m_i)^T\) are the within-class covariance matrixes,
- and \(S_W = S_1 + S_2\) is the total within-class covariance matrix

- Thus Fisher criterion and Generalize Rayleigh quotient are equivalent

\[
J(w) = \frac{(\bar{m}_1 - \bar{m}_2)^2}{\hat{s}_1^2 + \hat{s}_2^2} = \frac{w^T S_B w}{w^T S_W w}
\]
Solving the Two-Class Problem

\[ \max_w J(w) = \frac{w^T S_B w}{w^T S_W w} \]

- Note that the objective \( J \) is invariant w.r.t. to scaling \( w \rightarrow \alpha w \)
- Hence we can always choose \( w \) such that the denominator is simply \( w^T S_W w = 1 \)

- The constraint optimization problem follows
  \[ \max_w w^T S_B w, \quad s.t. \quad w^T S_W w = 1 \]

- The corresponding lagrangian is
  \[ L(w, \lambda) = w^T S_B w - \lambda (w^T S_W w - 1) \]
- \( \frac{\partial L}{\partial w} = 2S_B w - 2\lambda S_W w = 0 \)
- \( S_B w = \lambda S_W w \)

- The generalized eigenvalue problem follows
  \[ S_W^{-1} S_B w = \lambda w \]
- Observe that \( S_B w \) is always in the direction of \( m_1 - m_2 \), since \( \gamma = (m_1 - m_2)^T w \) is a scalar
  \[ S_B w = (m_1 - m_2)(m_1 - m_2)^T w = (m_1 - m_2)\gamma \propto (m_1 - m_2) \]
- Hence, we can always choose \( \lambda \) such that
  \[ w = S_W^{-1} (m_1 - m_2) \]
LDA from a Bayesian Standpoint

• Bayes Theorem

\[
Pr(\omega_k|x) = \frac{Pr(x|\omega_k) Pr(\omega_k)}{\sum_{i=1}^{K} Pr(x|\omega_i) Pr(\omega_i)}
\]

- \(Pr(\omega_k)\) the prior probability of class \(\omega_k\)
- \(Pr(x|\omega_k)\) the class-conditional density of \(x\) given class \(\omega_k\)

• Set class-conditional density of class \(\omega_k\) to be multivariate Gaussian

\[
Pr(x|\omega_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \exp\left[-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right]
\]

• LDA arises when the classes have a common covariance \(\Sigma_k = \Sigma \ \forall k\)

- In comparing two classes, it is sufficient to look at log-odds

\[
\log \frac{Pr(\omega_k|x)}{Pr(\omega_l|x)} = \log \frac{Pr(\omega_k)}{Pr(\omega_l)} - \frac{1}{2} (\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + x^T \Sigma^{-1} (\mu_k - \mu_l)
\]

- This equation is linear in \(x\)
  - The equal \(\Sigma\) cause canceling out the normalization factor and the quadratic terms in the exponent
LDA with Feature Independence

**Gaussian Naïve Bayes**

- Assume: \( \Pr(x|\omega_k) = \prod_{j=1}^{d} \Pr(x_j|\omega_k) \) where \( \Pr(x_j|\omega_k) \sim \mathcal{N}(x_j|\mu_{j,k}, \sigma_j^2) \)
  - Feature independence
  - Mean of \( x_j \) depends on class
  - Variance of \( x_j \) independent of class

- Then, for binary classification
  \[
  \ln \frac{\Pr(x_j|\omega_1)}{\Pr(x_j|\omega_2)} = \frac{(x_j - \mu_{j,2})^2}{2\sigma_j^2} - \frac{(x_j - \mu_{j,1})^2}{2\sigma_j^2} = \frac{\mu_{j,1} - \mu_{j,2}}{\sigma_j} \cdot \frac{1}{\sigma_j} \left( \frac{\mu_{j,1} + \mu_{j,2}}{2} - x_j \right)
  \]
  
  Distance between means
  Distance of \( x_j \) to midway point

- And the decision rule for Gaussians with class independent variances
  \[
  \ln \frac{\Pr(\omega_1|x)}{\Pr(\omega_2|x)} \propto \ln \frac{\Pr(\omega_1)}{\Pr(\omega_2)} - x^T\Sigma^{-1}(\mu_1 - \mu_2)
  \]
  \( \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2) \) is the diagonal covariance matrix

- This is linear decision boundaries \( f_{\mathbf{w}}(\mathbf{x}) = b + \mathbf{x}^T\mathbf{w} \)
LDA as a Linear Discriminant Function

• An equivalent description of the decision rule
  - Define linear discriminant function for class k
    \[ \delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \Pr(\omega_k) \]
  - Classify to the class with the largest value for its \( \delta_k(x) \)
    \[ g^*(x) = \arg \max_{\omega_k \in \Omega} \delta_k(x) \]

• Parameter Estimation
  - In practice, we do not know the parameters of the Gaussian distributions, and we will estimate them using a training set
    - \( \Pr(\omega_k) = \frac{n_k}{n} \), where \( n_k \) is the number of the class \( k \) observations
    - \( \hat{\mu}_k = \sum_{i \in \omega_k} x_i / n_k \)
    - \( \hat{\Sigma} = \sum_{k=1}^K \sum_{i \in \omega_k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K) \)
Quadratic Discriminant Analysis (QDA)

- If the covariance matrices $\Sigma_k$ are not equal, then the convenient cancellations (at the log-odds equation) do not occur
  - In particular the quadratic terms in $x$ remain

- This leads to *Quadratic Discriminant* functions

  $$\delta_k(x) = -\frac{1}{2} \log|\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log \Pr(\omega_k)$$

  - Parameter estimates are similar to LDA
    - but each class has a separate covariance matrices

- The number of parameters
  - For LDA, there are $(K - 1)(d + 1) + \frac{d(d+1)}{2}$ parameters
  - For QDA, there are $(K - 1)\left(\frac{d(d+1)}{2} + d + 1\right)$ parameters

- LDA and QDA both work really well
  - This is not because the data is necessarily Gaussian
  - Rather, for simple decision boundaries, Gaussian estimates are stable
Logistic Regression

The Two-Class Case

- The optimal decisions are based on posterior class probabilities \( \Pr(\omega_k|\mathbf{x}) \)
- For binary classification problems, the decisions are
  \[
  \log \frac{\Pr(\omega_1|\mathbf{x})}{1 - \Pr(\omega_1|\mathbf{x})} \geq 0
  \]
- Constrain the logit transformation of \( \Pr(\omega_1|\mathbf{x}) \) to be of a linear form
  \[
  \log \frac{\Pr(\omega_1|\mathbf{x})}{1 - \Pr(\omega_1|\mathbf{x})} = w_0 + \mathbf{x}^T \mathbf{w}
  \]
- This results with a specific form for the posterior class probability
  - The logistic model
    \[
    \Pr(\omega_1|\mathbf{x}) = \frac{1}{1 + \exp[-(w_0 + \mathbf{x}^T \mathbf{w})]}
    \]
    - Where \( g(z) = (1 + \exp[-z])^{-1} \) is the logistic (sigmoid) function
- Learning is aimed at estimating \( \mathbf{w} \) and \( w_0 \)
  - Rather than \( \Pr(\omega_k|\mathbf{x}) \)
Logistic Regression and Gaussian Distribution

- Logistic Regression is a Gaussian *discriminative* model with class independent covariance

\[
\log \frac{\Pr(\omega_1|x)}{\Pr(\omega_2|x)} = \log \frac{\Pr(\omega_1)}{\Pr(\omega_2)} - \frac{1}{2} (\mu_1 + \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) + x^T \Sigma^{-1} (\mu_1 - \mu_2) \\
= w_0 + x^T w
\]

- where
  - \( w_0 = \log \frac{\Pr(\omega_1)}{\Pr(\omega_2)} - \frac{1}{2} (\mu_1 + \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) \)
  - \( w = \Sigma^{-1} (\mu_1 - \mu_2) \)

- Learning is aimed at estimating \( w \) and \( w_0 \)
  - Rather than \( \Pr(\omega_1), \Pr(\omega_2), \mu_1, \mu_2, \Sigma \)

(*) *LDA is a Gaussian generative model with class independent covariance*
K-Class Logistic Regression

- The model is specified in terms of $K-1$ log-odds or logit transformations (reflecting the constraint that the probabilities sum to one)

$$\log \frac{\Pr(\omega_l|x)}{\Pr(\omega_K|x)} = \beta_{l0} + \beta_l^T x, \quad l = 1, \ldots, K - 1$$

- The choice of denominator is arbitrary, typically last class

- A simple calculation* shows that

  - $\Pr(\omega_l|x) = \frac{\exp(\beta_{l0}+\beta_l^T x)}{1+\sum_{j=1}^{K-1} \exp(\beta_{j0}+\beta_j^T x)}, \quad l = 1, \ldots, K - 1$

  - $\Pr(\omega_K|x) = \frac{1}{1+\sum_{j=1}^{K-1} \exp(\beta_{j0}+\beta_j^T x)}$

(* complete @ home)
Logistic Regression or LDA?

• Both are using the same linear form for the logit transformation

• LDA (Generative model)
  • Assumes Gaussian class-conditional densities and a common covariance
  • Model parameters are estimated by maximizing the full log likelihood
  \[
  \Pr(x, \omega_k) = \Pr(x) \Pr(\omega_k|x)
  \]
  • Parameters for each class are estimated independently of other classes
  • Easier to train, low variance, more efficient if model is correct
  • Higher asymptotic error, but converges faster

• Logistic Regression (Discriminative model)
  • Assumes class-conditional densities are of the (same) exponential family distribution
  • Model parameters are estimated by maximizing the conditional log likelihood
  \[
  \Pr(\omega_k|x)
  \]
  • Parameters are estimated while simultaneously considering all other classes
  • Harder to train, robust to uncertainty about the data generation process
  • Lower asymptotic error, but converges more slowly

• If normality holds, LDA is up to 30% more efficient; Otherwise logistic regression can be more robust. But the methods are similar in practice
Multiple Discriminant Analysis

- Multiple Discriminant Analysis is the generalization of Fisher Discriminant Analysis
  - Denote $C$ the number of classes and assume $C < d$, the dimension of the input space
  - Define $(C - 1)$ discriminant functions: $y_c = w_c^T x$, $c = 1, ..., C$
    - $y_c$ can be viewed as components or linear features of the $(C - 1)$ dimensional vector $y$
    - $\{w_c\}$ can be viewed as the columns of the matrix $W \in \mathbb{R}^{d \times (C - 1)}$
  - The projection from a $d$ to $(C - 1)$ dimensional space
    \[ y = W^T x \]

- Analogues to the two-class case, we seek a transformation matrix that will maximize the ratio of between-class and within-class scatter in the projected space
  - A simple scalar is the determinant of the scatter matrix
    - The determinant is the product of the eigenvalues and hence the product of the “variances” in the principle directions
    - It measures the square of the hyper-ellipsoidal scattering volume
Multiple Discriminant Analysis, Cont.

• The transformation matrix $W$ can be found by maximizing the following criterion

$$ J(W) = \frac{|\tilde{S}_B|}{|\tilde{S}_W|} = \frac{|W^T S_B W|}{|W^T S_W W|} $$

• $\tilde{S}_B, \tilde{S}_W$ are the projected scatter defined similarly to the between- and within-covariance matrixes
  - Where $\tilde{S}_W = \sum_c \sum_{y \in y_c} (y - \bar{m}_c)(y - \bar{m}_c)^T$ and $\tilde{S}_B = \sum_c (\bar{m}_c - \bar{m})(\bar{m}_c - \bar{m})^T$
    - $\bar{m}_c = \frac{1}{n_c} \sum_{y \in y_c} y$ and $\bar{m} = \frac{1}{n} \sum_c n_c \bar{m}_c$
    - It can be shown that $\tilde{S}_B = W^T S_B W$ and $\tilde{S}_W = W^T S_W W$

• It turns out that the columns of the optimal $W$ are the $(C - 1)$ generalized eigenvectors which corresponds to the largest eigenvalues in

$$ S_B w_c = \lambda_c S_W w_c $$

• Note that no more than $(C - 1)$ of the eigenvalue are nonzero
  - … because $S_B$ is the sum of $C$ rank 1 matrixes, but only $(C - 1)$ of them are independent (since $m = \frac{1}{n} \sum_c n_c m_c$)
More than Two Classes

- The proper generalization of the between-class covariance is not obvious
  - Duda & Hart (1973) suggested to derive it from the definition of the total covariance (of the whole data)
    
    \[
    S_T = \sum_{x \in X} (x - m)(x - m)^T \\
    = \sum_c \sum_{x \in X_c} (x - m_c + m_c - m)(x - m_c + m_c - m)^T \\
    = \sum_c \sum_{x \in X_c} (x - m_c)(x - m_c)^T + \sum_c \sum_{x \in X_c} (m_c - m)(m_c - m)^T 
    \]

- The generalized definitions follow naturally
  - Within-class covariance: \( S_W = \sum_c S_c = \sum_c \sum_{x \in X_c} (x - m_c)(x - m_c)^T \)
  - Between-class covariance: \( S_B = \sum_c n_c (m_c - m)(m_c - m)^T \)

- The total covariance is the sum of the within- and between-class covariance matrices
  
  \[ S_T = S_W + S_B \]
Most Discriminant Features (MDF)

- The LDA solution is given by the eigenvectors of the matrix $S_W^{-1}S_B$
- In practice, $S_W$ may be singular, e.g. when the data have large dimensionality while the size of the data set is much smaller ($M \ll N$)
- To alleviate this problem, we can perform two projections

  - Apply PCA to reduce dimensionality
  - Apply LDA to further reduce dimensionality and increase discriminatory representation

  - This two-steps approach is termed **Most Discriminant Features (MDF)**
Regularized Discriminant Analysis (RDA)

- A compromise between LDA and QDA
  - Shrink the separate covariance of QDA towards a common covariance as in LDA
    \[ \hat{\Sigma}_k(\alpha) = \alpha \hat{\Sigma}_k + (1 - \alpha) \hat{\Sigma} \]
  - Where \( \hat{\Sigma} \) is the pooled covariance matrix as used in LDA, and \( \alpha \in [0,1] \)
  - This method is very similar in flavor to Ridge regression

Figure 4.4: A two-dimensional plot of the vowel training data. There are eleven classes with \( X \in \mathbb{R}^{10} \), and this is the best view in terms of a LDA model (Section 4.3.3). The heavy circles are the projected mean vectors for each class. The class overlap is considerable.

Figure 4.7: Test and training errors for the vowel data, using regularized discriminant analysis with a series of values of \( \alpha \in [0,1] \). The optimum for the test data occurs around \( \alpha = 0.9 \), close to quadratic discriminant analysis.