Linear Methods for Classification

- Linear and Logistic Regression, LDA, QDA,
- k-NN (k Nearest Neighbors)
- optimal separating hyperplane – will be later (SVM)

Some Figures from Elem. of Stat. Learning (advanced book), the rest from Introduction to SL.
Classification

• We have a qualitative (categorical) goal variable $G$.
• The goal: classify to the true class $g$ from $G$.
• Often: probability $P(G=g \mid X)$ is predicted.
• Regression can be used,
• LOGISTIC regression is preferred over linear.
• Alternatives:
  • LDA linear discriminant analysis
  • k-NN k-nearest neighbours
  • SVM, decision trees and derived methods.
Example: Default Dataset

- **Goal:**
  - Will individual default on his/her payment?

- **Data:** <Income, Balance, Student, G=Default>

- Often displayed as color map.

- Only a fraction of non-default individual depicted.

- Individuals with default tend to have higher balances.
Remark: Notches

- We are 95% sure medians differ.
- We are not 95% individuum with default has higher balance than individuum without default.
Why Not Linear Regression?

- Really bad approach is to code diagnosis numerically
  - (since there is no ordering no scale).

\[ Y = \begin{cases} 
1 & \text{if stroke;} \\
2 & \text{if drug overdose;} \\
3 & \text{if epileptic seizure} 
\end{cases} \]

- Different coding could lead to very different model.
- If G has natural ordering
  - AND the gaps between values are similar
    the coding 1,2,3 would be reasonable.
Binary Goal Variable

- Coding 0/1 or -1/1 is possible.
- Still, logistic regression is preferred
  - no masking, no negative probabilities.

**FIGURE 4.2.** Classification using the Default data. Left: Estimated probability of default using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for default (No or Yes). Right: Predicted probabilities of default using logistic regression. All probabilities lie between 0 and 1.
Masking in Linear Regression for $G$

- We have three dummy variables Green, Blue, Orange, linear regression for each $P(g_i/x)$.

- or even linear cuts are possible.

better model:
Logistic Regression

- **logit function** \( \log[p/(1-p)] \),

- We create linear model for **transformed** input

\[
\log \frac{\Pr(G = 1|X = x)}{\Pr(G = 2|X = x)} = \beta_0 + \beta^T x.
\]

- The 'inverse' is called **logistic function**

\[
Pr(G = 1|X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}
\]
Fitting the Regression Coefficients

• We search **maximum likelihood** coefficients.

• Likelihood function:

\[
\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))
\]

• where \( p(X) = \Pr(Y = 1|X) \)

• **Probability** of the DATA given the model is called **likelihood** of the MODEL given the data.
(log) Likelihood

<table>
<thead>
<tr>
<th>Train Data</th>
<th>Predict</th>
<th>likelihood</th>
<th>loglik</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>G</td>
<td>Pgreen</td>
<td>Pblue</td>
</tr>
<tr>
<td>1</td>
<td>green</td>
<td>1/2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>green</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>3</td>
<td>blue</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>blue</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>1</td>
<td>yellow</td>
<td>1/2</td>
<td>0</td>
</tr>
</tbody>
</table>

-2-2log3

Logistic regression predicts:

\[
\Pr(G = k|X = x) = \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell0} + \beta_{\ell}^T x)}, \quad k = 1, \ldots, K - 1, \quad (4.18)
\]

Log-likelihood for \( N \) observations is

\[
\ell(\theta) = \sum_{i=1}^{N} \log p_{g_i}(x_i; \theta),
\]

where \( p_k(x_i; \theta) = \Pr(G = k|X = x_i; \theta) \).
Fitted Model

\[
\text{fit.g=glm(default~balance, family='binomial',data=Wage)}
\]

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. error</th>
<th>Z-statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-10.6513</td>
<td>0.3612</td>
<td>-29.5</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>balance</td>
<td>0.0055</td>
<td>0.0002</td>
<td>24.9</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

\[
\log \frac{\Pr(G = 1|X = x)}{\Pr(G = 2|X = x)} = \beta_0 + \beta^T x.
\]

**TABLE 4.1.** For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance. A one-unit increase in balance is associated with an increase in the log odds of default by 0.0055 units.

- **therefore**
  \[
P(\text{default }| \text{balance}) = \frac{e^{-10.6513 + 0.0055 \text{ balance}}}{1 + e^{-10.6513 + 0.0055 \text{ balance}}}
  \]
  \[
P(\neg \text{default }| \text{balance}) = \frac{1}{1 + e^{-10.6513 + 0.0055 \text{ balance}}}
  \]

- **generally**
  \[
  \Pr(G = k|X = x) = \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell0} + \beta_{\ell}^T x)}, \quad k = 1, \ldots, K - 1,
  \]
  \[
  \Pr(G = K|X = x) = \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell0} + \beta_{\ell}^T x)},
  \]  (4.18)
Discrete X: Coding (Automatic)

- Each except one value has its dummy variable.

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. error</th>
<th>Z-statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.5041</td>
<td>0.0707</td>
<td>-49.55</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>student[Yes]</td>
<td>0.4049</td>
<td>0.1150</td>
<td>3.52</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

**TABLE 4.2.** For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using student status. Student status is encoded as a dummy variable, with a value of 1 for a student and a value of 0 for a non-student, and represented by the variable student[Yes] in the table.

\[
\hat{Pr}(\text{default}=\text{Yes}|\text{student}=\text{Yes}) = \frac{e^{-3.5041+0.4049\times1}}{1+e^{-3.5041+0.4049\times1}} = 0.0431
\]

\[
\hat{Pr}(\text{default}=\text{Yes}|\text{student}=\text{No}) = \frac{e^{-3.5041+0.4049\times0}}{1+e^{-3.5041+0.4049\times0}} = 0.0292
\]
Multiple Logistic Regression

\[
\log \left( \frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,
\]

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. error</th>
<th>Z-statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-10.8690</td>
<td>0.4923</td>
<td>-22.08</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>balance</td>
<td>0.0057</td>
<td>0.0002</td>
<td>24.74</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>income</td>
<td>0.0030</td>
<td>0.0082</td>
<td>0.37</td>
<td>0.7115</td>
</tr>
<tr>
<td>student[Yes]</td>
<td>-0.6468</td>
<td>0.2362</td>
<td>-2.74</td>
<td>0.0062</td>
</tr>
</tbody>
</table>

**TABLE 4.3.** For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance, income, and student status. Student status is encoded as a dummy variable student[Yes], with a value of 1 for a student and a value of 0 for a non-student. In fitting this model, income was measured in thousands of dollars.
More Response Classes

• Model has the form:

\[ \log \frac{\Pr(G = 1 | X = x)}{\Pr(G = K | X = x)} = \beta_{10} + \beta_1^T x \]
\[ \log \frac{\Pr(G = 2 | X = x)}{\Pr(G = K | X = x)} = \beta_{20} + \beta_2^T x \]
\[ \vdots \]
\[ \log \frac{\Pr(G = K - 1 | X = x)}{\Pr(G = K | X = x)} = \beta_{(K-1)0} + \beta_{K-1}^T x. \]

• To probabilities:

\[ \Pr(G = k | X = x) = \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell0} + \beta_{\ell}^T x)}, \quad k = 1, \ldots, K - 1, \]
\[ \Pr(G = K | X = x) = \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell0} + \beta_{\ell}^T x)} \]
Confounding

**FIGURE 4.3.** Confounding in the Default data. Left: Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of balance, while the horizontal broken lines display the overall default rates. Right: Boxplots of balance for students (orange) and non-students (blue) are shown.

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. error</th>
<th>Z-statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-10.8690</td>
<td>0.4923</td>
<td>-22.08</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>balance</td>
<td>0.0057</td>
<td>0.0002</td>
<td>24.74</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>income</td>
<td>0.0030</td>
<td>0.0082</td>
<td>0.37</td>
<td>0.7115</td>
</tr>
<tr>
<td>student[Yes]</td>
<td>-0.6468</td>
<td>0.2362</td>
<td>-2.74</td>
<td>0.0062</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. error</th>
<th>Z-statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.5041</td>
<td>0.0707</td>
<td>-49.55</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>student[Yes]</td>
<td>0.4049</td>
<td>0.1150</td>
<td>3.52</td>
<td>0.0004</td>
</tr>
</tbody>
</table>
LDA - Linear Discriminant Analysis

- assumes Normal distribution $X$ for each class $q$,

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$
Bayes' Theorem

- Prior probabilities of classes: \( \pi_k \).
- Probability of \( x \) given the class model \( k \):

\[
f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)
\]

- Posterior probability:

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

\[
= \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma}} \exp \left( -\frac{1}{2\sigma^2} (x - \mu_k)^2 \right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma}} \exp \left( -\frac{1}{2\sigma^2} (x - \mu_l)^2 \right)}.
\]
Example:

FIGURE 4.4. Left: Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary. Right: 20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.

• If prior probability of green class is lower, the decision boundary moves to the left.
Bayes Boundary

- Assume we know the true distribution of data.
- For each $x$, predict the class $g_j$ with the highest $P(G=g_j|X=x)$.
- No better prediction can be made.
- The error of such model is called **Bayes error** this gives lower bound for our classifiers.
LDA – One Dimension

• We calculate from the data:

$$\hat{\pi}_k = \frac{n_k}{n}$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\Sigma = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K)$$

• We predict the class with maximal:

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

It comes from the logarithm of probabilities, terms in all deltas are erased.
LDA – Multiple Dimensions

• We calculate from the data:

\[ \hat{\pi}_k = \frac{n_k}{n} \]

\[ \hat{\mu}_k = \frac{1}{n_k} \sum_{i : y_i = k} x_i \]

\[ \hat{\sigma}^2 = \frac{1}{n - K} \sum_{k = 1}^{K} \sum_{i : y_i = k} (x_i - \hat{\mu}_k)^2 \]

• We predict the class with maximal:

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

It comes from the logarithm of probabilities, terms in all deltas are erased.
Confusion matrix
Evaluation of a Classifier

<table>
<thead>
<tr>
<th>Predicted default status</th>
<th>True default status</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
<td>9,644</td>
<td>252</td>
</tr>
<tr>
<td>Yes</td>
<td>23</td>
<td>81</td>
</tr>
<tr>
<td>Total</td>
<td>9,667</td>
<td>333</td>
</tr>
</tbody>
</table>

TABLE 4.4. A confusion matrix compares the LDA predictions to the true default statuses for the 10,000 training observations in the Default data set. Elements on the diagonal of the matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were misclassified. LDA made incorrect predictions for 23 individuals who did not default and for 252 individuals who did default.

- Classification error: \( \frac{252+23}{10000} = 0.0275 \)
- Is this classifier:
  - almost perfect
  - slightly better then trivial
  - bad?
Different Error Costs

\[
\begin{array}{|c|cc|c|}
\hline
\text{Predicted default status} & \text{True default status} & \\
\text{No} & \text{Yes} & \text{Total} \\
\hline
\text{No} & 9,644 & 252 & 9,896 \\
\text{Yes} & 23 & 81 & 104 \\
\hline
\text{Total} & 9,667 & 333 & 10,000 \\
\hline
\end{array}
\]

- Default occurred 333, we recognized 81 – it is we missed 252 defaults.
- We can consider as risky all with \( p > 0.2 \), then we miss less defaults.
  - green: total error
  - blue: 'default=Y' error
  - orange: non-default
ROC Curve

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>+ or Non-null</th>
<th>- or Null</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>True class</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- or Null</td>
<td>False Pos. (FP)</td>
<td>True Neg. (TN)</td>
<td>N</td>
</tr>
<tr>
<td>+ or Non-null</td>
<td>True Pos. (TP)</td>
<td>False Neg. (FN)</td>
<td>P</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Pos. rate</td>
<td>FP/N</td>
<td>Type I error, 1—Specificity</td>
</tr>
<tr>
<td>True Pos. rate</td>
<td>TP/P</td>
<td>1—Type II error, power, sensitivity, recall</td>
</tr>
<tr>
<td>Pos. Pred. value</td>
<td>TP/P*</td>
<td>Precision, 1—false discovery proportion</td>
</tr>
<tr>
<td>Neg. Pred. value</td>
<td>TN/N*</td>
<td></td>
</tr>
</tbody>
</table>
QDA – k Covariance Matrices

„Elipses may be different for each class“.  
• More parameters in the model.  
• Both LDA and QDA are often used.

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

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

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

- Error Cost matrix $L$ dim $K \times K$, $K$ number of $g$ in $G$.
- 0 on the diagonal, non-negative everywhere $L(k,l)$ cost of predict true $G_k$ to be $G_l$.

\[
\begin{align*}
EPE &= E[L(G, \hat{G}(X))] , \\
EPE &= E_X \sum_{k=1}^{K} L[G_k, \hat{G}(X)]Pr(G_k | X) \\
\hat{G}(x) &= \arg\min_{g \in G} \sum_{k=1}^{K} L(G_k, g)Pr(G_k | X = x).
\end{align*}
\]

With the 0–1 loss function this simplifies to

\[
\hat{G}(x) = \arg\min_{g \in G} [1 - Pr(g | X = x)]
\]

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

- Bayes classifier, bayes rate.
QDA or Basis Expansion usually not a big difference

**FIGURE 4.6.** Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in Figure 4.1 (obtained using LDA in the five-dimensional space $X_1, X_2, X_1X_2, X_1^2, X_2^2$). The right plot shows the quadratic decision boundaries found by QDA. The differences are small, as is usually the case.
Comparison of Classifiers

- LDA – assumes normal distribution,
- logistic regression assumes less,
- both leads to linear decision boundary.

gaussian true dist.  corellated x  t-distribution (more flat)
Comparison 2

- if assumptions are met – better prediction with fewer data,
- assumptions not met – often worst prediction.
Summary

• Linear regression only for two-valued goal G.
• LDA, if we assume two normal distributed classes (it is more stable),
• logistic regression – usually similar to LDA,
• QDA – sometimes may be useful,
• k-NN can capture any non-linear decision boundary. For simple boundaries may give worst predictions.
Ahead:
Optimal Separating Hyperplane

**Figure 4.14.** A toy example with two classes separable by a hyperplane. The orange line is the least squares solution, which misclassifies one of the training points. Also shown are two blue separating hyperplanes found by the perceptron learning algorithm with different random starts.
FIGURE 4.9. Although the line joining the centroids defines the direction of greatest centroid spread, the projected data overlap because of the covariance (left panel). The discriminant direction minimizes this overlap for Gaussian data (right panel).