# Data Science II
# (P8106)

Department of Biostatistics
Mailman School of Public Health
Columbia University

# Classification

- Qualitative response: take values in an unordered set $C$
  - email $\in$ {spam, ham}
  - disease status $\in$ {diseased, undiseased}
- A feature vector $X$
- A qualitative response $Y$ taking values in the set $C$
- Task: build a function $C(X)$ that takes $X$ as input and predict the value for $Y$, i.e. $C(X) \in C$
- Often we are more interested in estimating the probabilities that $Y$ belongs to each category in $C$

# Classification

- Training data: $\{(x_1, y_1), \ldots, (x_n, y_n)\}$
- $y_1, \ldots, y_n$ are qualitative
- Training error: $\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$, $\hat{y}_i$ is the predicted class label
- Test observation of the form $(x_0, y_0)$
- Test error: $\text{Ave}(I(y_0 \neq \hat{y}_0))$
- The test error is minimized, on average, by a very simple classifier that assigns each observation to the most likely class, given its predictor values

# The Bayes Classifier

- Assign a test observation with predictor vector $x_0$ to the class $j$ for which $Pr(Y = j \mid X = x_0)$ is largest
- Example: $C = \{\text{blue}, \text{orange}\}$, predict the class label to be orange if $Pr(Y = \text{orange} \mid X = x_0) > 0.5$
- Bayes decision boundary

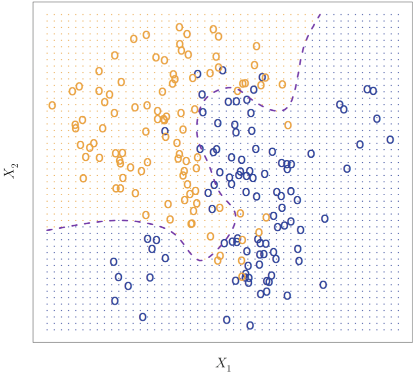



Figure: ISL 2.13

# Bayes error rate

- The Bayes classifier produces the lowest possible test error rate - the Bayes error rate
- The error rate at $X = x_0$

$$1 - \max_j Pr(Y = j \mid X = x_0)$$

- Overall Bayes error rate

$$1 - E\left(\max_j Pr(Y = j \mid X)\right)$$

# Evaluating classification models

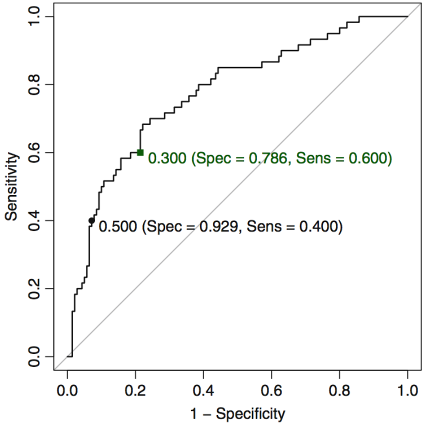
- Binary outcome, $p(X) = Pr(Y = 1|X)$
- Rule: Predict $Y = 1$ if $\hat{p}(X) > c$
- Fix $c$

| | Observed $Y = 1$ | Observed $Y = 0$ |
|---|---|---|
| Predicted $Y = 1$ | TP | FP |
| Predicted $Y = 0$ | FN | TN |

- Sensitivity = TP/(TP+FN)
- Specificity = TN/(FP+TN)
- Receiver Operating Characteristic (ROC) Curves
  - Plot true positive rate (sensitivity) versus false positive rate (1-specificity) with varying $c$
  - A higher ROC curve is more favorable

# ROC curve

- The best possible prediction method would yield the point (0,1), representing 100% sensitivity and 100% specificity
- A random guess would give a point along a diagonal line
- Area under the curve (AUC)



# Linear methods for classification

- Logistic Regression
- Linear discriminate analysis

# Can we use linear regression?

- Since $E(Y|X = x) = Pr(Y = 1|X = x)$, we might consider regression
- Perform a linear regression of $Y$ on $X$ and classify as Yes if $\hat{Y} > 0.5$ ?
- Linear regression might produce probabilities less than zero or larger than one
- Logistic regression is more appropriate

# Logistic regression

- Write $p(X) = Pr(Y = 1|X)$
- Logistic regression uses the form

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- Apply *logit* transformation

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

- Use maximum likelihood to estimate the parameters
- Prediction:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X}}$$

- Decision boundary?

# Linear vs. logistic regression

- Orange: observed data (Y vs. X)
- Blue: fitted value of $p(X)$
- With logistic regression, $\hat{p}(X)$ is in $[0, 1]$

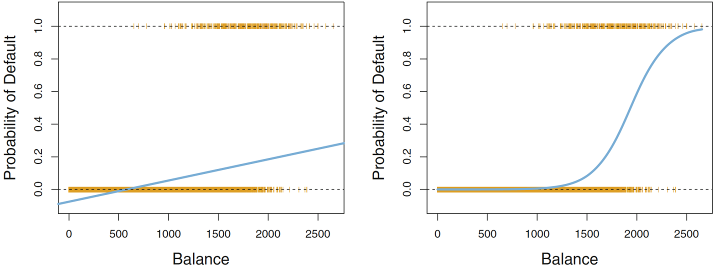



Figure: ISL 4.2

# Logistic regression with several variables

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}$$

# Penalization

- Maximize

$$\log \ell(\beta) - \lambda \left\{ (1-\alpha)\frac{1}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j| \right\}$$

- $\lambda$ controls the total amount of penalization
- $\alpha$ is the “mixing proportion”
- `glmnet(...,family="binomial")`

# GAM for classification

- Model

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(X_1) + \ldots + f_p(X_p)$$

- `gam(..., family = binomial)`