Overview

Methods for classification

- logistic regression
- discriminant analysis (linear & quadratic) and naive Bayes classifier
- k-nearest neighbor classifier

Logistic regression

Outcome variables may not be continuous variables that follow a normal distribution

- dichotomous data: presence or absense of a symptom

  $\rightarrow$ linear regression is still a good classifier in this case (i.e., link with linear discriminant analysis) although predicted probabilities are not between 0 and 1

- polytomous data (orderer or unordered): eye color or educational level

- proportion of times invited for a job interview

- count data: number of infections during a month

- time before an event happens (e.g., employment, cancer)

Linear regression is not a good option here !

Logistic regression

Generalized linear models: 3 components

- linear predictor (with/without random effects)

  $b_0 + b_1(months)_i$

  $\rightarrow$ in general, no level 1-error term

- outcome variable with a specific (error) distribution with a certain mean (and variance)

  $(stress)_i \sim Bernoulli(\pi_i)$

  $\rightarrow$ π_i denotes the mean (e.g., a proportion when the outcome is Bernoulli)

- a link function

  $logit(\pi_i) = log(\frac{\pi_i}{1-\pi_i})$

Logistic regression

$$
\begin{aligned}
(mathematics)_i &\sim Bernoulli(\pi_i) \\
logit(\pi_i) &= log(\frac{\pi_i}{1-\pi_i}) = b_0 + b_1(studytime)_i + b_2(schoolsup)_i
\end{aligned}
$$

```
> m_log = glm( G1bin ~ studytime + schoolsup , data=student , family=binomial )

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.03865    0.31006   0.125 0.900808
studytime     0.56435    0.15421   3.660 0.000253 ***
schoolsupyes -1.26865    0.31698  -4.002 6.27e-05 ***
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 466.64  on 393  degrees of freedom
Residual deviance: 437.78  on 391  degrees of freedom
AIC: 443.78

Number of Fisher Scoring iterations: 4
```

Logistic regression

- is based on maximizing a likelihood (Fisher scoring)
  - maximizes the probability (likelihood) of the observed data given the parameters
  - likelihood depends on $(p+1)$-dimensional vector of parameters $\boldsymbol{\beta}$ (p is the number of predictors)

- Fisher scoring algorithm
  - iterative procedure for estimating $\boldsymbol{\beta}^{(t)}$ $(t = 0, 1, \dots)$
  - based on gradient $S(\boldsymbol{\beta}^{(t)})$ and expected Hessian $\mathcal{H}(\boldsymbol{\beta}^{(t)})$ from the likelihood evaluated at $\hat{\boldsymbol{\beta}}^{(t)}$
  - $\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - \mathcal{H}(\hat{\boldsymbol{\beta}}^{(t)})^{-1} S(\hat{\boldsymbol{\beta}}^{(t)})$
    - $\rightarrow$ gradient: vector of first-order partial derivatives of a function in terms of its parameters
    - $\rightarrow$ Hessian: matrix of second-order partial derivatives (i.e., derivative of the first derivative)

Logistic regression

- is based on maximizing a likelihood (Fisher scoring)
- you have to specify the *family* (is related to the distribution of the dependent variable)

  $\rightarrow$ *binomial* is standard choice for dichotomous/binary data

- no level 1-error (and associated variance term σ_e^2)
- watch out for overdispersion (scale factor)
- differences for interpretation
  - regression coefficient β: effect of predictor on logit scale (log odds)
  - e^{β}: effect of predictor on the odds (issues of multicollinearity)

    $\rightarrow$ per unit increase in the predictor, the odds is multiplied with e^{β}

    $\rightarrow$ logistic curve between predictor and (predicted probability for the) outcome

Discriminant analysis

Discriminant analysis

- similar to logistic regression but a different point of view
- logistic regression: focus on conditional distribution of the outcome given the predictors
  - $logit(\pi_i) = log(\frac{\pi_i}{1-\pi_i}) = b_0 + b_1(studytime)_i + b_2(schoolsup)_i$
  - $E[p(Y_i = k|X = x_i)] = \pi_i = \frac{exp^{b_0+b_1(studytime)_i+b_2(schoolsup)_i}}{1+exp^{b_0+b_1(studytime)_i+b_2(schoolsup)_i}}$
- linear discriminant analysis: focus on joint distribution of outcome and predictors
  - modeling the density/distribution of the predictors in each outcome class (i.e., p-variate normal distribution with μ_k and Σ)
  - use marginal/prior probability of outcome: $p(Y = k)$ (i.e., class size)
  - use Bayes theorem to get the conditional distribution of the outcome given the predictors $\rightarrow$ classification is based on this conditional distribution

Discriminant analysis

Discriminant analysis

- Linear Discriminant Analysis (LDA)
  - classification based on the discriminant score (i.e., score for each observation for each class)
  - is a linear function of the predictors: linear decision boundaries
- other view on LDA
  - discriminant variables as linear combinations of the variables

    $\rightarrow$ weights are based on an eigenvalue analysis
  - with maximal separation of the classes (visualisation in 2D)
- Quadratic Discriminant Analysis (QDA)
  - class-specific means and covariance matrix: μ_k and Σ_k
  - discriminant scores are now a quadratic function of the predictors: quadratic decision boundaries

Naive Bayes classifier

Naive Bayes classifier

- conditional independence assumption within a class (Σ_k diagonal)
- gives biased estimates of the probabilities (due to naive assumptions)
- but can give good classification: you only need to know which probability is the largest
  - increases bias: not very flexible and misspecified model
  - serious decrease in variance (especially for large p): not many parameters

K-Nearest Neighbor

K-Nearest Neighbor (KNN) classifier

- determining the smoothness of the decision boundaries
  - small K: very flexible boundaries
    $\rightarrow$ What happens if $K = 1$?
  - large K: very smooth (but less flexible) boundaries
    $\rightarrow$ goes in the direction of linear decision boundaries
    $\rightarrow$ what happens if $K = N$?

Classification (evaluation of classifiers)

Evaluation of a classifier

- do not look at training error (too optimistic)
- confusion matrix: total error rate and compare with null-rate (i.e., assign all cases to largest outcome class)
- two types of error
  - false positive and false negative rate (i.e., specificity and sensitivity)
  - vary threshold for classification: ROC-curve
  - best classifier has largest Area Under the Curve (AUC)

Classification (bias-variance trade off)

The bias-variance trade off

- bias
  - a more flexible model (i.e., QDA versus LDA or KNN with a large K) will reduce bias because it makes less stringent assumptions on decision boundaries (closer to "true" model)
- variance: overfitting
  - a more flexible model may increase variance because it has more parameters that need to be estimated (to make the method more flexible)
  - ratio of number of cases for each parameter is important

Classification (bias-variance trade off)

The bias-variance trade off

- when the complexity of a model grows, in general, there is a tendency for
  - in the beginning, decrease in bias is stronger than increase in variance
  - later on, the (increase in) variance starts to outweigh the (decrease in) bias
  - look for an optimal model more in the middle
  - always consider parsimony: when two models perform almost the same, take the least flexible (i.e., the simpler) one

    $\rightarrow$ because you only have an estimate for classifier performance (e.g., cross-validation estimate for test error), you never know the true performance in reality

Classification (comparison)

Which method to choose for classification?

- linear regression: never
- logistic regression
  - parametric: linear decision boundaries (quite strict)
  - boundaries can be made quadratic when adding a quadratic term
  - problematic with small n and large p
  - not easy to work with when the outcome has multiple categories (i.e., various generalizations exist)
  - breaks down when classes are well separated

Classification (comparison)

Which method to choose for classification?

- linear discriminant analysis (LDA)
  - parametric: assumption of linear boundaries (is quite strict) and assumption of normality and equal covariance matrix within a class

    $\rightarrow$ also in logistic regression (but different way of estimating the parameters)
  - may work well for small n (because you want to reduce variance in that case)
  - works well when classes are well separated and when there are more than 2 classes
  - problematic for large p (too many parameters to estimate)

Classification (comparison)

Which method to choose for classification?

- quadratic discriminant analysis (QDA)
  - parametric: assumption of quadratic decision boundaries (is more flexible) and assumption of normality within a class
  - covariance matrix per class: many parameters need to be estimated (bad for the variance)
  - may work for small n (but not too small because many parameters have to be estimated)
  - can be considered in between LDA and KNN with small K
- naive Bayes (NB) classifier
  - makes assumptions that are probably wrong ("naive")
  - only method that may work well in high-dimensional (large p) case

Classification (comparison)

Which method to choose for classification?

- K-nearest neighbor
  - non-parametric: no assumption on the shape of the decision boundaries (good for very non-linear problems)
  - performance heavily depends on optimal choice for smoothness of the decision boundary (i.e., value of K; problem of overfitting)
  - bad performance for small n (LDA/QDA probably better)

    $\rightarrow$ assumptions on the shape of the boundaries needed in that case
  - does not tell which predictors are important in the classification (i.e., no regression coefficients or discriminant weights)

- other options: support vector machines (chapter 9), trees-boosting-bagging-random forest (Chapter 8), neural networks, kernel methods

Classification (exercises)

Performing various classification methods in R

Exercise (student data set from last session)

```
> student_full = read.csv2( "student-mat.csv" )
> student_full$G1bin = student_full$G1>8 # to make a dichotomous outcome
> student = student_full[ c("G1","G1bin","studytime","absences",
"freetime","failures","age") ]
```

- use the variable *G1bin* which indicates whether a student passed (score of 8 or higher) or failed the test as outcome and the selected variables (*studytime*, *absences*, *freetime*, *failures*, *age*) as predictors
- fit a logistic regression model with the five main effects and estimate the test error rate (misclassification rate) by means of 10-fold cross-validation
- estimate the test error rate (by using the validation set approach) for this model when using linear and quadratic discriminant analysis and the naive Bayes classifier. Which method yields the best results?

Classification (exercises)

Exercise (student data set from last session)

- Choose the optimal value of K (try values for K going from 1 up to 50) for the KNN classifier (by means of CV). Determine the test error (by means of the validation approach) of the optimal KNN model. Use the 5 main effects for the KNN classifier.

- extra: choose 4 nested models and fit logistic regression, LDA, QDA and KNN to each of these models. Determine the best model in terms of complexity with cross-validation. Use the validation approach to determine the test error for logistic regression, LDA, QDA and KNN. Which of these methods works best for this data set?
  - note that you have to determine the optimal K for KNN (use CV to determine the optimal K).