

Statistics for Affymetrix GeneChips

Comparing Multiple Groups; Discrimination Analysis

<http://www.isrec.isb-sib.ch/~darlene/NCCR-PS/>

PLANT SURVIVAL
National Centre of Excellence in Research

NCCR Plant Survival, 19-20 March 2007 Lec 4

PLANT SURVIVAL
National Centre of Excellence in Research

NCCR Plant Survival, 19-20 March 2007 Lec 4

Linear models

- Linear in *parameters*
- Simplest version: comparing single treatment (T) to single control (C)

$$Y_C = \mu + \varepsilon_C; \hat{\mu} = Y_C$$

$$Y_T = \mu + \alpha + \varepsilon_T; \hat{\alpha} = Y_T - Y_C$$
- With multiple observations, the estimates are averages (or differences of averages)
- (We just saw an example of this)
- Readily extends *to more conditions*

PLANT SURVIVAL
National Centre of Excellence in Research

NCCR Plant Survival, 19-20 March 2007 Lec 4

Design Matrix and Contrasts

- The *design matrix* indicates the hybs (which RNA hybridized to each array)
- In **limmaGUI** and **affy1mGUI** you can find the design matrix by clicking on 'Advanced'
- The *contrasts* are the comparisons of interest
- In **limmaGUI** and **affy1mGUI** you set the contrasts by specifying the comparisons you are interested in

PLANT SURVIVAL
National Centre of Excellence in Research

NCCR Plant Survival, 19-20 March 2007 Lec 4

Advantages of linear models

- The linear modeling approach might look more complicated, but there are *several advantages*:
 - Linear modeling is already well-established in statistics
 - Analyze *all arrays together* combining information in optimal way
 - Combined estimation of *precision*
 - *Extensible* to arbitrarily complicated experiments
- We can use linear modeling to compare *multiple groups*

PLANT SURVIVAL
National Centre of Excellence in Research

NCCR Plant Survival, 19-20 March 2007 Lec 4

ANOVA

- The idea:

PLANT SURVIVAL
National Centre of Excellence in Research

NCCR Plant Survival, 19-20 March 2007 Lec 4

A two-group comparison

- Comparison between 2 groups: *wild type* (WT) and *mutant* (MT) - interest is in the difference MT - WT
- Say we have 5 chips: 2 WT and 3 MT

Chip	Target
Chip 1	WT
Chip 2	WT
Chip 3	MT
Chip 4	MT
Chip 5	MT



NCCR Plant Survival, 19-20 March 2007

Lec 4

Experimental vs. Observational studies

- Controlled experiment**: subjects assigned to groups by the investigator
 - randomization*: protects against bias in assignment to groups
 - blind, double-blind*: protects against bias in outcome assessment/measurement
 - placebo*: fake 'treatment'
- Observational study**: subjects 'assign' themselves to groups
 - confounder*: associated with both group membership and the outcome of interest



NCCR Plant Survival, 19-20 March 2007

Lec 4

Observational studies

- Advantages**
 - often *easier* to carry out
 - don't 'interfere' with the system, what you see is '*natural*' rather than 'artificial'
 - variation is *biologically relevant*, as it has been unaltered
 - sometimes manipulation is *not possible*
- Drawbacks**
 - confounders
 - association (correlation) is NOT causation



NCCR Plant Survival, 19-20 March 2007

Lec 4

Factorial crossing

- Compare 2 (or more) sets of conditions in the *same experiment* - more efficient than separate single factor experiments
- Designs with factorial treatment structure allow you to measure *interaction* between two (or more) sets of conditions that influence the response
- Like other experiments, factorial designs may be either *observational* or *experimental*



NCCR Plant Survival, 19-20 March 2007

Lec 4

3 types of 2-factor factorial designs

- 2 experimental factors - you *randomize* treatments to each unit
- 2 observational factors - you *cross-classify* your populations into groups and get a sample from each population
- 1 experimental and 1 observational factor - you *get a sample* of units from each population, *then use randomization* to assign levels of the experimental factor (treatments), separately within each sample



NCCR Plant Survival, 19-20 March 2007

Lec 4

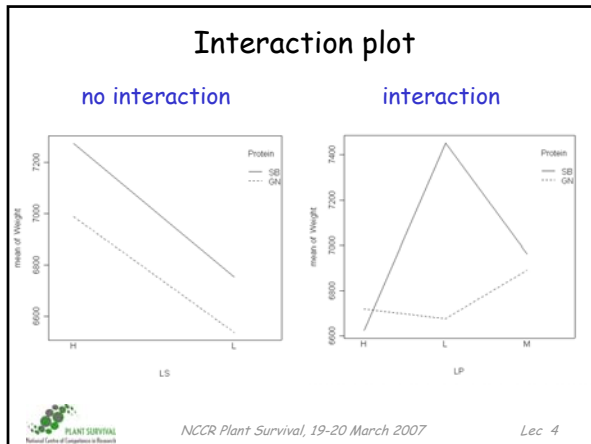
Interaction

- Interaction is very common (and very important) in science
- Interaction is a *difference of differences*
- Interaction is present if the effect of one factor *is different* for different levels of the other factor
- Main effects can be difficult to interpret in the presence of interaction*, because the effect of one factor depends on the level of the other factor



NCCR Plant Survival, 19-20 March 2007

Lec 4



Example: Setup

- Continuing on from the earlier example, suppose that cells extracted from the WT and MT individuals are either *Stimulated* (S) or *Unstimulated* (U)
- Treatment conditions for the 5 chips:

Chip	Strain	Treatment
Chip 1	WT	U
Chip 2	WT	S
Chip 3	MT	U
Chip 4	MT	S
Chip 5	MT	S

PLANT SURVIVAL
National Centre of Genomics & Research

NCCR Plant Survival, 19-20 March 2007 Lec 4

Example: Factorial structure

- There are 2 factors here: *Strain* and *Treatment*
- Each factor has 2 levels, and all 4 combinations of Strain and Treatment are present:
 - WT.U, WT.S, MT.U, MT.S (which is replicated)
- So, this a 2x2 factorial design

PLANT SURVIVAL
National Centre of Genomics & Research

NCCR Plant Survival, 19-20 March 2007 Lec 4

Example: Questions of interest

- For this example, assume that the comparisons of interest are which genes respond:
 - to stimulation in WT (WT.S - WT.U)
 - to stimulation in MT (MT.S - MT.U)
 - differently* in MT compared to WT (MT.S - MT.U) - (WT.S - WT.U)
- Question 3 concerns a difference of differences, i.e. is addressing *interaction*

PLANT SURVIVAL
National Centre of Genomics & Research

NCCR Plant Survival, 19-20 March 2007 Lec 4

Estrogen experiment

- In the practical this afternoon, you will continue analyzing the estrogen experiment that you started yesterday
- This is an example of an experiment with a factorial design
- You can look at the most differentially expressed genes between single conditions or sets of conditions

PLANT SURVIVAL
National Centre of Genomics & Research

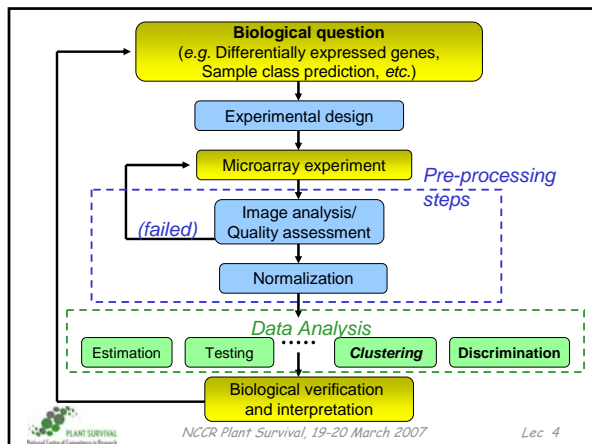
NCCR Plant Survival, 19-20 March 2007 Lec 4

Summary

- The linear modeling approach is powerful and flexible
- For many types of experiments, it is simple to apply
- It becomes easier with practice 😊
- You will get some more practice this afternoon!

PLANT SURVIVAL
National Centre of Genomics & Research

NCCR Plant Survival, 19-20 March 2007 Lec 4



Classification

- Historically, *objects* are classified into *groups*
 - periodic table of the elements (chemistry)
 - taxonomy (zoology, botany)
- Why classify?
 - organizational convenience, convenient summary
 - prediction
 - explanation
- *Note:* these aims do not necessarily lead to the same classification; e.g. *SIZE* of object in hardware store vs. *TYPE/USE* of object

Classification, cont

- Classification divides objects into groups based on a set of values
- Unlike a theory, a classification is neither true nor false, and should be judged largely on the usefulness of results (Everitt)
- However, a classification (clustering) may be useful for suggesting a theory, which could then be tested

Classification

- *Task:* assign objects to classes (groups) on the basis of measurements made on the objects
- *Supervised:* classes are predefined, want to use a (training or learning) set of labeled objects to form a classifier for classification of future observations (discrimination analysis)
- *Unsupervised:* classes unknown, want to discover them from the data (cluster analysis)

Discrimination

- Objects (e.g. arrays) are to be classified as belonging to one of a number of *predefined classes* {1, 2, ..., K}
- Each object associated with a class label (or *response*) $Y \in \{1, 2, \dots, K\}$ and a feature vector (vector of predictor variables) of G measurements: $\mathbf{X} = (X_1, \dots, X_G)$
- *Aim:* predict Y from \mathbf{X}

Classifiers

- A *predictor* or *classifier* partitions (divides) the space of gene expression profiles into K disjoint subsets, A_1, \dots, A_K , such that for a sample with expression profile $\mathbf{X}=(X_1, \dots, X_G)$ in A_k the predicted class is k
- Classifiers are built from a *learning set (LS)*

$$L = (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$$
- *Classifier* C built from a learning set L :

$$C(\cdot, L): \mathbf{X} \rightarrow \{1, 2, \dots, K\}$$
- *Predicted class* for observation \mathbf{X} :

$$C(\mathbf{X}, L) = k \text{ if } \mathbf{X} \text{ is in } A_k$$

Fisher Linear Discriminant Analysis

First applied in 1935 by M. Barnard at the suggestion of R. A. Fisher (1936), *Fisher linear discriminant analysis (FLDA)*:

1. finds linear combinations of the gene expression profiles $X = X_1, \dots, X_G$ with large ratios of between-groups to within-groups sums of squares - *discriminant variables*;
2. predicts the class of an observation X by the class whose mean vector is closest to X in terms of the discriminant variables



NCCR Plant Survival, 19-20 March 2007

Lec 4

Nearest Neighbor Classification

- Based on a measure of *distance* between observations (e.g. Euclidean distance)
- k-nearest neighbor rule (Fix and Hodges (1951)) classifies an observation X as follows:
 - find the k observations in the learning set *closest* to X
 - predict the class of X by *majority vote*, i.e., choose the class that is most common among those k observations
- The number of neighbors k can be chosen by *cross-validation* (more on this later)



NCCR Plant Survival, 19-20 March 2007

Lec 4

Classification Trees

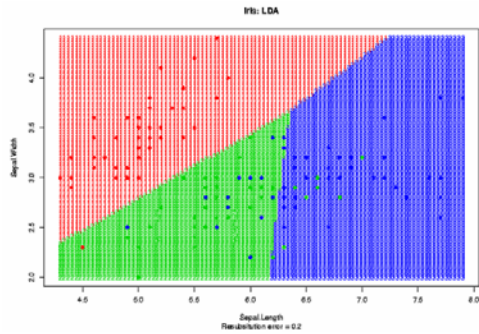
- Partition the feature space into a set of rectangles, then fit a simple model in each one
- Binary tree structured classifiers* are constructed by repeated splits of subsets (nodes) of the measurement space X into two descendant subsets (starting with X itself)
- Each terminal subset is assigned a class label; the resulting partition of X corresponds to the classifier



NCCR Plant Survival, 19-20 March 2007

Lec 4

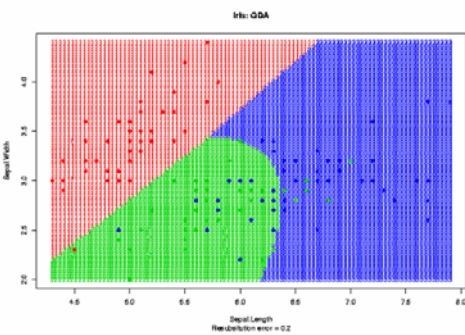
Example: linear discriminant analysis



NCCR Plant Survival, 19-20 March 2007

Lec 4

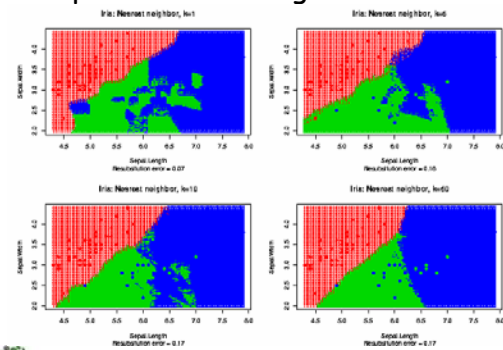
Example: quadratic discriminant analysis



NCCR Plant Survival, 19-20 March 2007

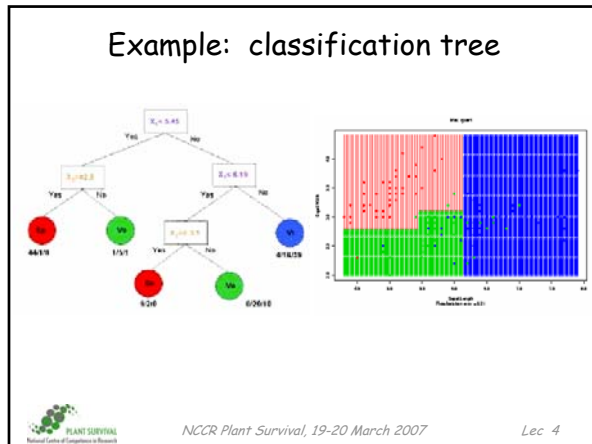
Lec 4

Example: nearest neighbor classifier



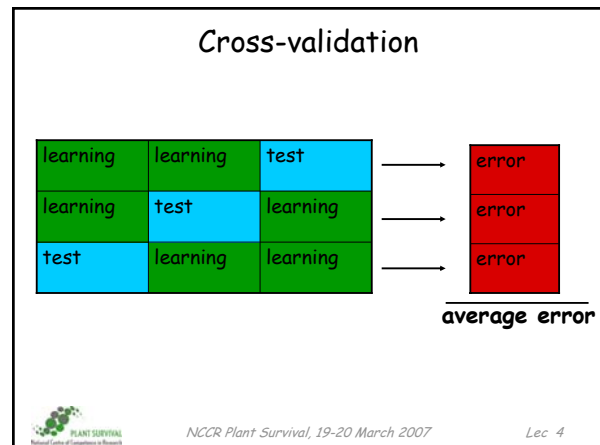
NCCR Plant Survival, 19-20 March 2007

Lec 4



- ### R: discrimination
- A number of R packages contain functions to carry out discrimination, including:
 - MASS: lda, qda
 - sma: dlda
 - class: knn
 - rpart: classification and regression trees (recursive partitioning)
 - ipred: bagging
 - e1071: svm
 - LogitBoost: boosting
 - randomForest: trees with bagging
- NCCR Plant Survival, 19-20 March 2007 Lec 4

- ### Assessing classifier performance
- Resubstitution estimation:** error rate on the learning set
 - Problem: downward bias
 - Test set estimation:** divide cases in learning set into two sets, L_1 and L_2 ; classifier built using L_1 , error rate computed for L_2
 - Problem: reduced effective sample size
 - V-fold cross-validation (CV) estimation:** Cases in learning set randomly divided into V subsets of (nearly) equal size (for example, 5 or 10 subsets). Build classifiers leaving one set out; test set error rates computed on left out set and averaged to get overall error
 - Bias-variance tradeoff: smaller V can give larger bias but smaller variance
- NCCR Plant Survival, 19-20 March 2007 Lec 4



- ### More on performance assessment
- Common (**BUT WRONG**) to do feature selection using all of the data, then CV only for model building and classification
 - However, usually features are unknown and the intended inference includes feature selection => CV estimates as above tend to be *downward biased*
 - Features should be selected only *from the learning set* used to build the model (and not the entire set)
- NCCR Plant Survival, 19-20 March 2007 Lec 4

- ### How NOT to estimate error
- ** DON'T DO THIS ** DON'T DO THIS ****
 - Use the *whole data set* to choose which variables (features) to use in the classifier
 - Divide the data into (10, say) subsets for CV
 - Leave out a subset and build a classifier with features chosen from the *whole data set*
 - Use the classifier to predict the left out subset
 - Average over left out subsets to estimate error
 - ** DON'T DO THIS ** DON'T DO THIS ****
- NCCR Plant Survival, 19-20 March 2007 Lec 4

Summary

- Classification (clustering or discrimination) more appropriate for larger studies
- What we have seen here is an overview - there is still *much more* to classification!
- Often, people (wrongly) carry out clustering when it is more appropriate to perform a discrimination analysis (that is, when groups are actually 'known')
- If you do need to carry out classification tasks, it is best to collaborate with an *experienced analyst*

