Statistics for the Sciences

Classification

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- Introduction to Classification
- Multinomial Logistic Regression
- K-Nearest Neighbors (KNN)
- Linear Discriminant Analysis (LDA)
- Lab 1: KNN
- Lab 2: Multinomial Logistic Regression
- Lab 3: LDA

Introduction

- Classification involves predicting a **categorical/qualitative response**:
  - Cancer versus Normal
  - Tumor Type 1 versus Tumor Type 2 versus Tumor Type 3
- Classification problems tend to occur even more frequently than regression problems in biomedical applications.
- Categorical/qualitative variables take values in an unordered set: e.g.
  - eye color $\in$ {brown, blue, green}
- We want to build a function that $C(X)$ takes as input the feature vector $X = (X_1, \ldots, X_p)$ and predicts the value for Y, i.e. $C(X)$ is in which category.
- Often we are more interested in estimating the probability that X belongs to a given category.
  - For example: we might want to know the probability that someone will develop diabetes, rather than to predict whether or not they will develop diabetes.

Introduction

- Let n be the total number observations. For example, the data $(x1, y1), \ldots, (x_n, y_n)$. And let $\hat{y}$ be our estimate.
- Then the fitted **error rate**, the proportion of mistakes that are made is

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i).$$

Multinomial Logistic Regression

- Let $\mathcal{C}$ be the set of collection of responses of Y. For example, a specific cancer is one of $\mathcal{C}$ = (Cancer, Normal), digit class is one of $\mathcal{C}= \{0, 1, \ldots, 9\}$.
- Is there an ideal $C(X)$? Suppose the K elements in $\mathcal{C}$ are numbered $1, 2, \ldots, K$. For any x, let

$$p_k(x) = P(Y = k|X = x), k = 1, \ldots, K.$$

  These are the **conditional class probabilities** at $X = x$.
- Then the **Bayes Classifier** at x is

$$C(x) = j, \text{ if } p_j(x) = \max\{p_1(x), \ldots, p_K(x)\}$$

  For responses with two categories, we just check which probability is greater than 50%.

Multinomial Logistic Regression

- The error rate of the Bayes classifier at $X = x_0$ will be

$$1 - P(\text{Correct classification}|X = x_0) = 1 - \max_j Pr(Y = j|X = x_0).$$

  - This error rate represents the lowest possible error rate achievable by any classifier because the Bayes classifier uses the true conditional probabilities and always selects the class with the highest probability.
- In general, the overall Bayes error rate is given by

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

- We can build parametric models for representing the conditional class probabilities $p_k(x)$ to construct a Bayes classifier.
  - Logistic regression is such a method when $k = 2$.

Multinomial Logistic Regression

- Multinomial Logistic Regression is a generalization of logistic regression when the response has more than 2 classes/levels.
- To illustrate the idea, suppose there is only one predictor X and Y has K levels. Then

$$P(Y_i = j|x_i) = \frac{\exp(\alpha_j + x_i\beta_j)}{\sum_{h=1}^{K} \exp(\alpha_h + x_i\beta_h)}, j = 1, 2, \cdots, K.$$

- The constraint on the parameters is $\sum_{j=1}^{K} p_j = 1$
- With intercepts, there are $2(K-1)$ parameters in the above Multinomial Logistic Regression model.

K-Nearest Neighbors

- Can we take a totally non-parametric (model-free) approach to classification?
- K-nearest neighbors (KNN):
  - 1. For any given X_0, identify the k observations whose X values are *closest to the observation* X_0 *at which we want to make a prediction.*
  - 2. Classify the observation of interest X_0 to the most frequent class label of those K nearest neighbors: If the majority of the Y's are orange we predict orange otherwise guess blue.
- The smaller that k is the more flexible the method will be.

K-Nearest Neighbors

- Example: K-nearest neighbors in two dimensions

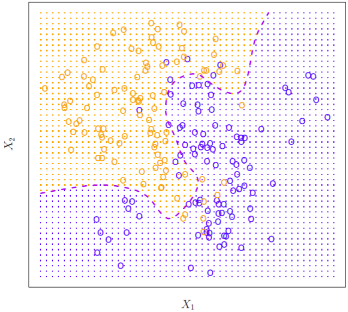
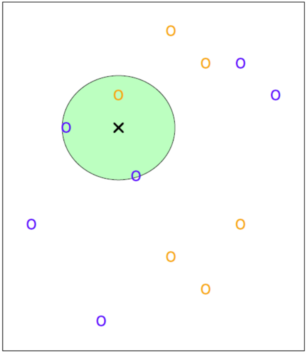
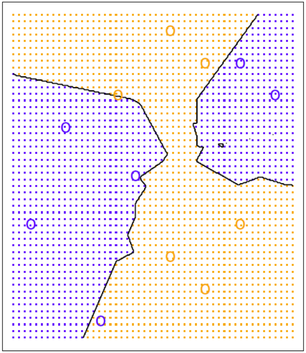


Figure 1: A simulated data set consisting of 100 observations in each of two groups, indicated in blue and in orange. The purple dashed line represents the Bayes decision boundary. The orange background grid indicates the region in which a test observation will be assigned to the orange class, and the blue background grid indicates the region in which a test observation will be assigned to the blue class.

K-Nearest Neighbors

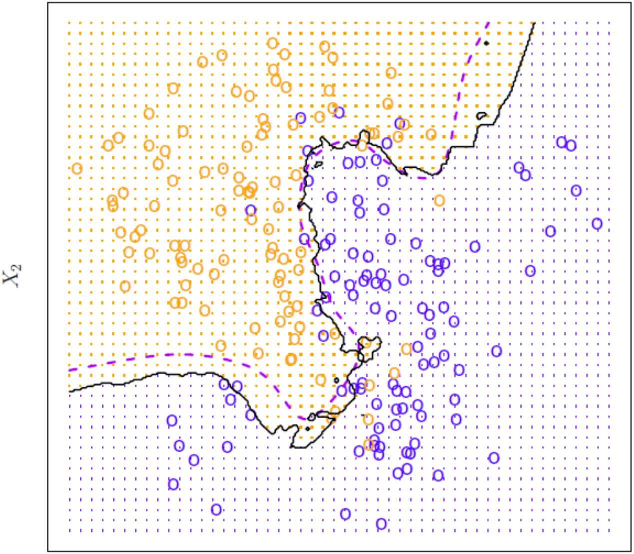
- Example: KNN Example with $k = 3$



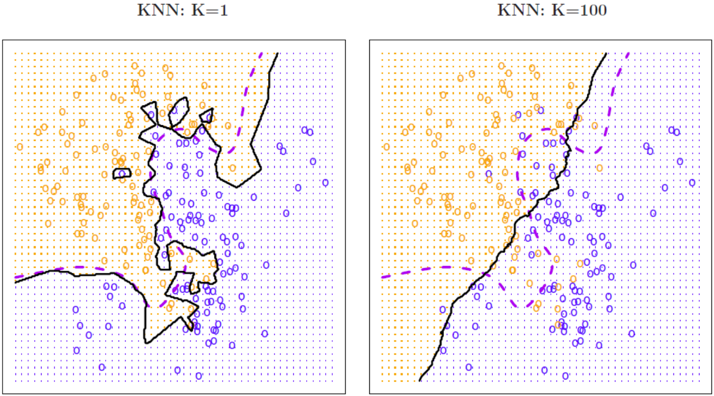


K-Nearest Neighbors

- Example: KNN Example with $k = 10$



K-Nearest Neighbors



K-Nearest Neighbors

- When k is small, the model is more flexible or complex.
- When k is large, the classification boundary will be smooth.
- We must always keep in mind that More flexible/complicated is not always better!
  - The ideal K can be chosen by Cross-Validation, a popular model selection method in machine learning.
- Advantages
  - Simple, intuitive, model-free.
  - Good option when p is very small.
- Curse of dimensionality: when p is large, no neighbours are "near". All observations are close to the boundary.

Linear Discriminant Analysis

- Suppose we now have information on $f_k(x) = Pr(X = x|Y = k)$, distribution of the predictors within each class,
  - How do we use this to make predictions?
- We apply the Bayes Rule in probability:

$$p_k(x) = Pr(Y = k|X = x) = \frac{Pr(X = x|Y = k) \cdot Pr(Y = k)}{Pr(X = x)}$$

  or

$$p_k(x) = Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)},$$

  where in the formula, we need
  - $f_k(x) = Pr(X = x|Y = k)$ which is the density for X in class k, $k = 1, \dots, K$
  - $\pi_k = Pr(Y = k), k = 1, \dots, K$ which is the marginal or prior probability of Y for class k.
- We refer to $p_k(x)$ as the posterior probability that an observation posterior $X = x$ belongs to the *kth* class.

Linear Discriminant Analysis

π_k is generally simple to estimate:

- If our data are a random sample of size n, then we can use the sample proportion

$$\hat{\pi}_k = \frac{\#\{Y = k\}}{n},$$

  which is the fraction of the training observations that belong to the kth class.
- Otherwise can use outside information (eg. historical data)

Linear Discriminant Analysis

- Estimate of $f_k(x) = Pr(X = x|Y = k)$ is more difficult. This is a **density estimation** problem.
  - Technically the notation $f_k(x) = Pr(X = x|Y = k)$ is only correct if X is a discrete random variable. If X is continuous, $f_k(x)dx$ would correspond to the probability of X falling in in a small region dx around x.
- In LDA (Linear Discriminant Analysis), we will use Gaussian/normal densities for these, separately in each class.

Linear Discriminant Analysis when $p = 1$

- There is only one predictor X.
- The Gaussian density has the form

$$\boxed{f_k(x) = \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}, -\infty < x < \infty.}$$

  Here μ_k is the mean, and σ_k^2 is the variance in class k, $k = 1, \ldots, K$.
  - We will assume that all the $\sigma_k = \sigma$ are the same.
- Plugging this into Bayes formula,

$$p_k(x) = \frac{\pi_k \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{i=1}^{K} \pi_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma}\right)^2}}, k = 1, \ldots, K.$$

- To classify at the value $X = x$, we need to see which of the $p_k(x)$ is largest.
- Happily, there are simplifications and cancellations.

Linear Discriminant Analysis when $p = 1$

- Taking logs, and discarding terms that do not depend on k, we see that this is equivalent to assigning x to the class with the largest **discriminant score**:

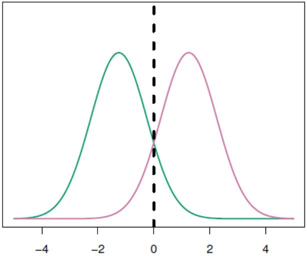
  $$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

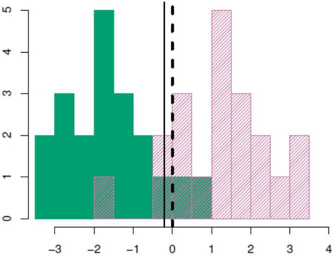
  which is a linear function of x.
- If there are $K = 2$ classes and $\pi_1 = \pi_2 = 0.5$, then one can show that the decision boundary is at

  $$x = \frac{\mu_1 + \mu_2}{2}.$$

Linear Discriminant Analysis when $p = 1$

- Example with $\mu_1 = -1.5, \mu_2 = 1.5$, $\pi_1 = \pi_2 = 0.5$, and $\sigma = 1$.





- Typically we don't know these parameters; we simply estimate the parameters and plug them into the rule.

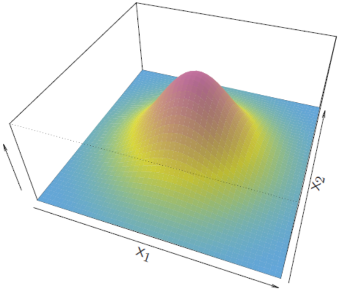
Linear Discriminant Analysis when $p = 1$

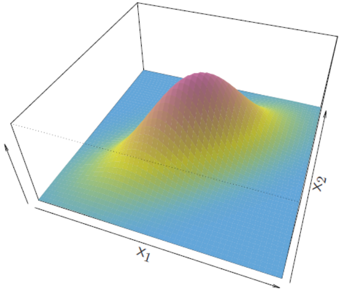
$$
\begin{aligned}
\hat{\pi}_k =& \frac{n_k}{n}, n_k \text{ is the number of observations in class } k \\
\hat{\mu}_k =& \frac{1}{n_k} \sum_{i:y_i=k} x_i \\
\hat{\sigma}^2 =& \frac{1}{n-k} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2 \\
=& \frac{1}{n_k} \sum_{k=1}^{K} (n_k - 1)\hat{\sigma}_k^2,
\end{aligned}
$$

where $\hat{\sigma}_k^2 = \dfrac{1}{n_k} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$ is the sample variance for the kth class. That is, $\hat{\sigma}^2$ is the pooled estimate of the common variance σ^2.

Linear Discriminant Analysis when $p > 1$

- When $p > 1$, we consider multivariate normal distribution for $f_k(x) = Pr(X = x|Y = k), k = 1, \ldots, K$.
- For example, bivariate normal density when $p = 2$



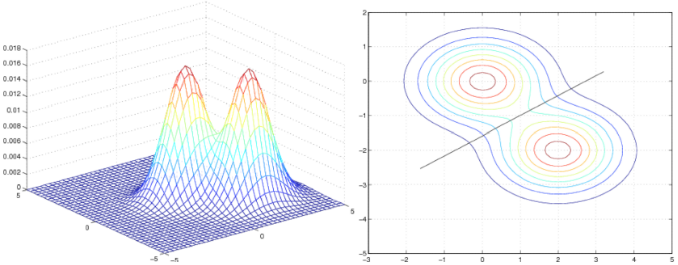


Linear Discriminant Analysis when $p > 1$

- Discriminant function:

$\delta_k(\mathbf{x}) = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log \pi_k$. Despite its complex form, it is a linear function of X:

$$\delta_k(\mathbf{x}) = c_{k0} + c_{k1}x_1 + \cdots + c_{kp}x_p$$



Linear Discriminant Analysis

- From $\delta_k(x)$ back to probabilities:

Once we have estimates $\delta_k(x)$, we can turn these into estimates for class probabilities:

$$\widehat{Pr}(Y = k|X = x) = \frac{e^{\delta_k(x)}}{\sum_{i=1}^{K} e^{\delta_i(x)}}$$

- So classifying to the largest $\delta_k(x)$ amounts to classifying to the class for which $\widehat{Pr}(Y = k|X = x)$ is largest.
- When $K = 2$, we classify to class 2 if $\widehat{Pr}(Y = 2|X = x) > 0.5$, else to class 1.

Lab

- Example (`feinmorph.csv`):https://mjkeough.github.io/examples/feinbergLDA.nb.html
  - Feinberg et al. (2014) examined morphological, genetic, and acoustic (call) criteria in four species of leopard frogs (Rana sphenocephala, R. pipiens, R. palustris, and a new species named R. kauffeldi) and an acoustically similar congener R. sylvatica (acoustic criteria only). For 283 museum specimens across the first four species, they measured size (snout-vent length) and 12 other morphological characteristics: head length, head width, eye diameter, tympanum diameter, foot length, eye to naris distance, naris to snout distance, thigh length, internarial distance, interorbital distance, shank length, and dorsal snout angle. Foot length was not recorded for 19 specimens, so these were excluded from the analysis. They also recorded seven call characteristics (call length, call rate, call rise time, call duty cycle, pulse number, pulse rate, and dominant frequency) from 45 frogs in the field across the five species. Call rate and call length were both adjusted based on regressions against temperature to a standard 14^{o}C.
  - Response spp with four classes: `"rk"`, `"rpa"`, `"rpi"`, `"rsph"`

Lab

```
## 'data.frame':    264 obs. of  14 variables:
##  $ spp: Factor w/ 4 levels "rk","rpa","rpi",..: 1 1 1 1 1 1 1 1 1
##  $ svl: num  67.7 61.4 53.7 68.2 71.7 ...
##  $ hw : num  22.4 21.3 19.2 22.4 23.4 ...
##  $ hl : num  21.6 19.8 17.6 21.7 21.2 ...
##  $ td : num  5.18 5.37 4.14 5.7 5.75 4.5 4.75 4.39 5.46 4.7 ...
##  $ ew : num  3.13 4.6 3.49 5.39 5.27 4.59 4.29 4.11 5.15 4.88 ...
##  $ tl : num  27.4 31.4 25.4 30.1 33.1 ...
##  $ sl : num  37.8 35.1 27.8 38.8 38.6 ...
##  $ fl : num  59.1 51.8 44.1 59.9 61.1 ...
##  $ end: num  4.83 4.19 3.45 4.71 4.45 4.13 4.36 3.67 3.68 3.19 ..
##  $ nsd: num  4.52 2.82 2.72 4.67 4.63 4.31 3.99 3.59 3.47 3.39 ..
##  $ iod: num  4.35 4.21 3.25 4.38 4.76 3.86 3.48 3.6 3.73 3.73 ...
##  $ ind: num  4.49 2.97 3.89 5.58 5.6 3.81 4.54 4.64 4.6 4.67 ...
##  $ dsa: num  1.09 1.13 1.15 1.09 1.17 ...
```

Lab 1

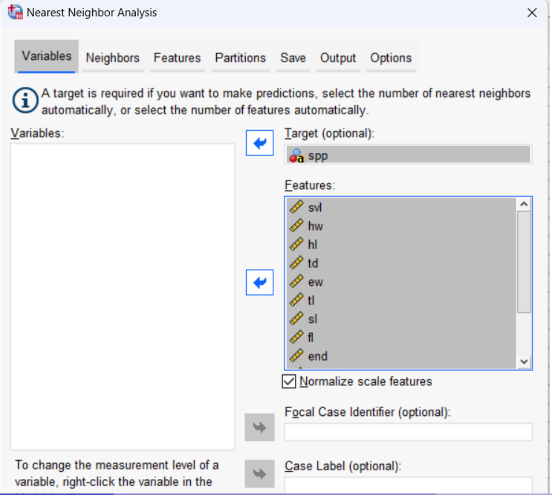
- KNN method with R

```
library(class)
y=feinmorph[,1]
X=feinmorph[,-1]
knn.pred=knn(X, X, y, k=10);
table(knn.pred, y);
```

```
##           y
## knn.pred   rk rpa rpi rsph
##     rk    152  21  14   24
##     rpa     3   8   1    3
##     rpi     0   1  11    0
##     rsph    2   0   5   19
```

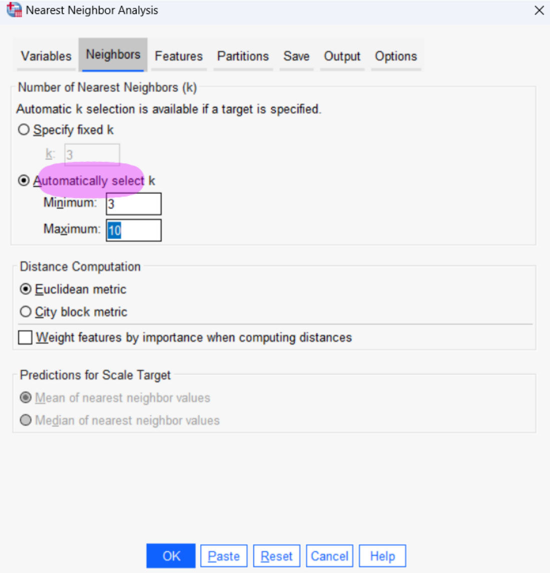
Lab 1

- After importing data `feinmorph.csv` to SPSS, click on `Analyze` → `Classify` → `Nearest Neighbor ...`, and then add the Dependent variable and Features (Covariates)
  - By default, the covariates will be standardized



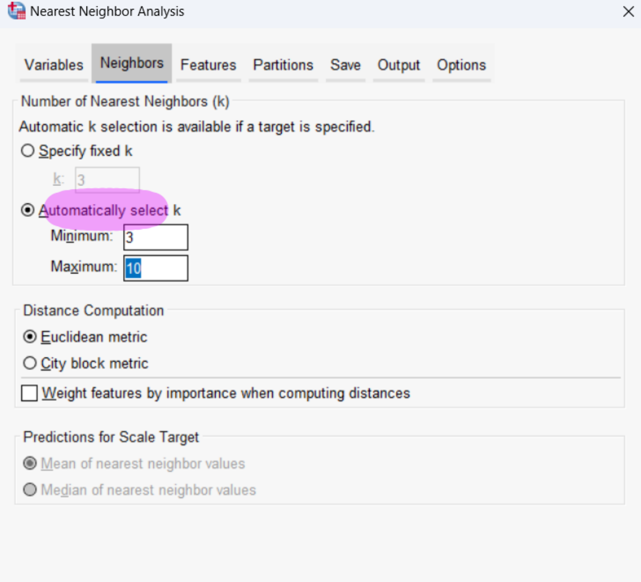
Lab 1

- By default, $k = 3$



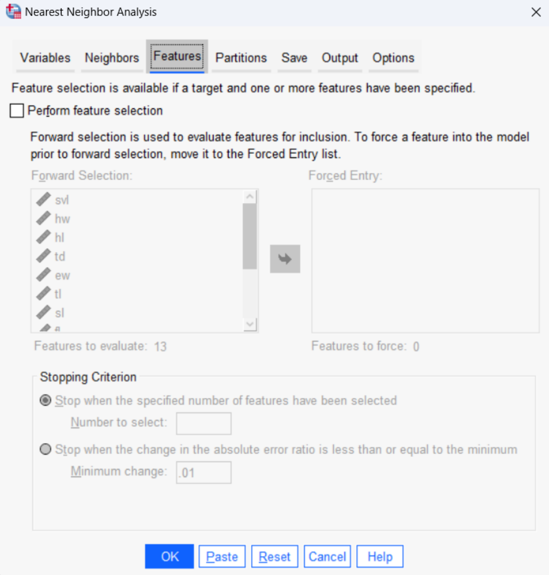
Lab 1

- By default, $k = 3$



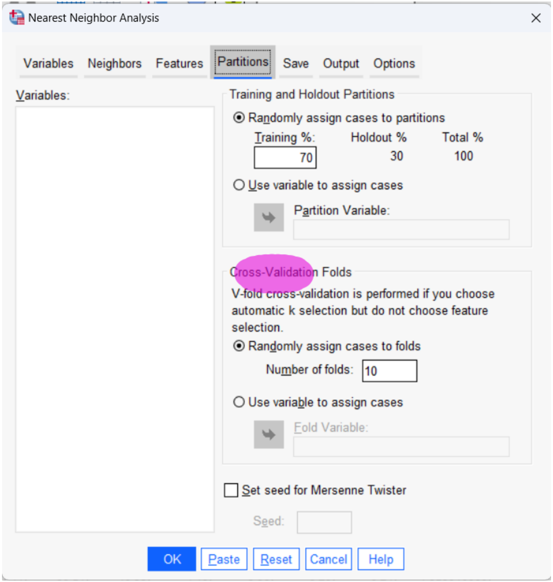
Lab 1

- Let's choose no model selection



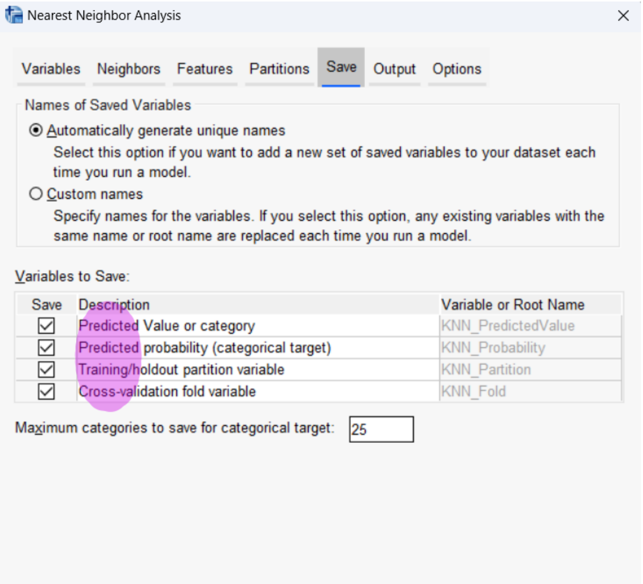
Lab 1

- Let's try using Cross-validation to conduct model selection (the best k value)

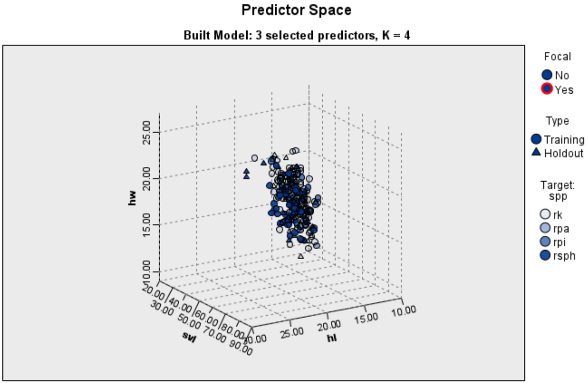


Lab 1

- Save some outputs



Lab 1

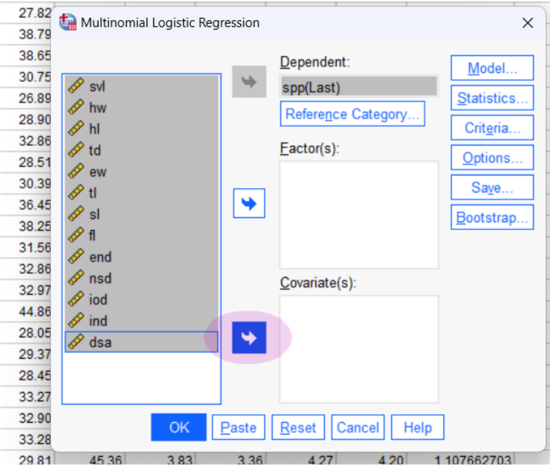


Select points to use as focal records

This chart is a lower-dimensional projection of the predictor space, which contains a total of 13 predictors.

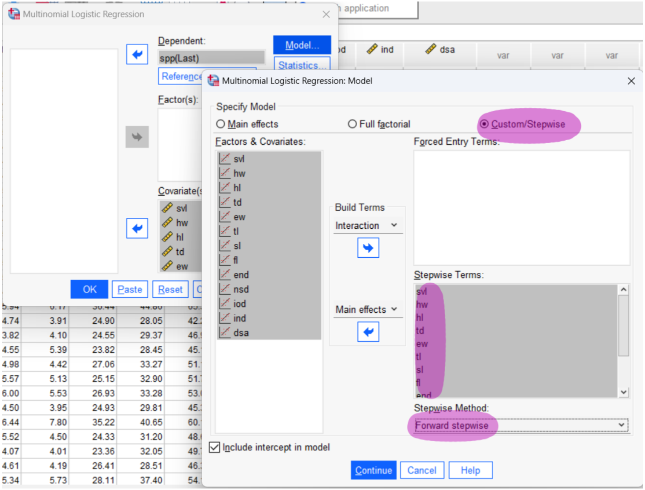
Lab 2

- We now fit a multinomial logistic regression.
- Click on `Analyze` → `Regression` → `Multinomial Logistic ...`, and then add the Dependent variable and Covariates



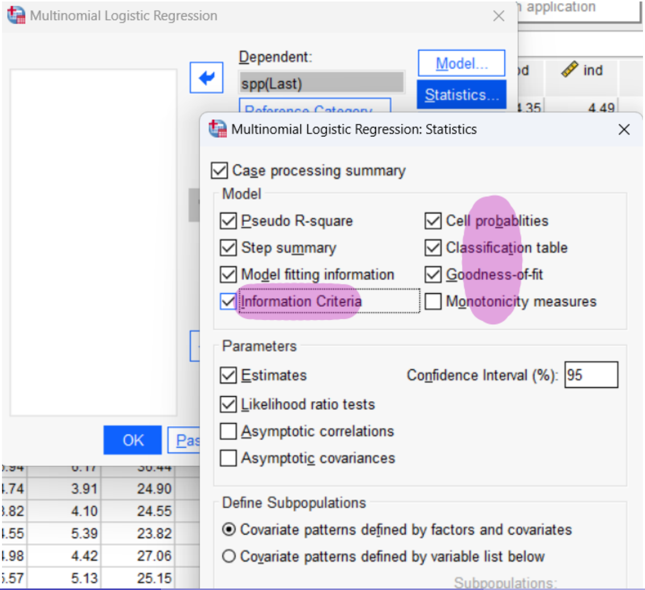
Lab 2

- Click on `Model...` and set up for model selection



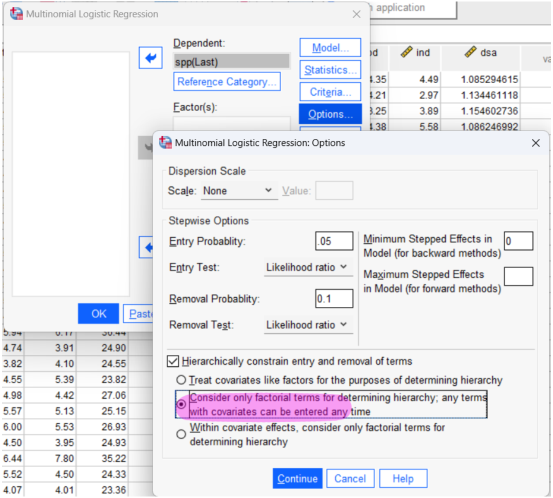
Lab 2

- Click on `Statistics...` to select additional statistics



Lab 2

- Click on `Options...`



Lab 2

- After setting up all options, click `OK` to run the multinomial logistic regression

Model Fitting Information

	Model Fitting Criteria			Likelihood Ratio Tests		
Model	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	593.226	603.954	587.226			
Final	306.590	424.596	240.590	346.636	30	<.001

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	1947.118	759	<.001
Deviance	240.590	759	1.000

Pseudo R-Square

Cox and Snell	.731
Nagelkerke	.820
McFadden	.590

Lab 2

Likelihood Ratio Tests

Effect	Model Fitting Criteria: AIC of Reduced Model	Model Fitting Criteria: BIC of Reduced Model	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	Likelihood Ratio Tests: df	Likelihood Ratio Tests: Sig.
Intercept	335.917	443.195	275.917	35.326	3	<.001
svl	313.522	420.800	253.522	12.932	3	.005
hw	315.749	423.028	255.749	15.159	3	.002
td	332.320	439.599	272.320	31.730	3	<.001
ew	325.945	433.223	265.945	25.355	3	<.001
sl	335.236	442.515	275.236	34.646	3	<.001
end	325.811	433.090	265.811	25.221	3	<.001
nsd	314.509	421.787	254.509	13.919	3	.003
iod	312.704	419.983	252.704	12.114	3	.007
ind	309.401	416.680	249.401	8.811	3	.032
dsa	341.195	448.474	281.195	40.605	3	<.001

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

Lab 2

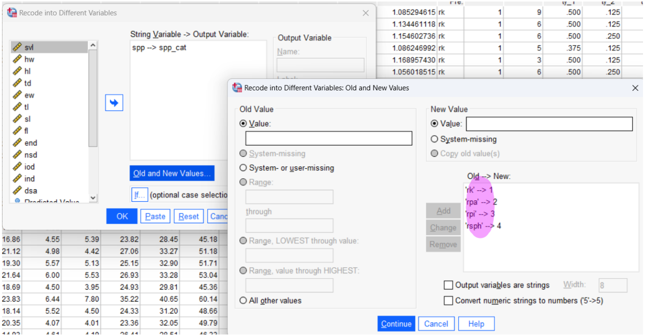
Parameter Estimates

spp[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
rk	Intercept	-6.082	3.809	2.550	1	.110			
	svl	-.137	.116	1.381	1	.240	.872	.695	1.096
	hw	.482	.383	1.584	1	.208	1.620	.764	3.431
	td	.637	.622	1.048	1	.306	1.891	.558	6.404
	ew	-.374	.272	1.881	1	.170	.688	.404	1.174
	sl	-.278	.189	2.171	1	.141	.757	.523	1.096
	end	-2.711	.639	17.973	1	<.001	.066	.019	.233
	nsd	-.215	.519	.172	1	.678	.806	.291	2.231
	iod	1.616	.598	7.307	1	.007	5.034	1.559	16.249
	ind	1.180	.560	4.442	1	.035	3.254	1.086	9.746
	dsa	15.866	4.981	10.146	1	.001	7770958.080	447.296	135006204651.4
rpa	Intercept	-11.004	4.923	4.995	1	.025			
	svl	-.388	.142	7.417	1	.006	.679	.514	.897
	hw	.178	.484	.135	1	.714	1.195	.462	3.086
	td	-1.915	.766	6.259	1	.012	.147	.033	.660
	ew	-.787	.407	3.733	1	.053	.455	.205	1.011
	sl	.666	.254	6.877	1	.009	1.946	1.183	3.200
	end	-2.610	.845	9.544	1	.002	.074	.014	.385
	nsd	-.164	.660	.062	1	.804	.849	.233	3.096
	iod	.231	.705	.107	1	.744	1.259	.317	5.010
	ind	1.934	.743	6.774	1	.009	6.914	1.612	29.658
	dsa	23.185	6.032	14.773	1	<.001	11719676304.28	86001.263	1.597E+15
rpi	Intercept	-33.344	7.404	20.281	1	<.001			
	svl	.089	.199	.198	1	.656	1.093	.740	1.614
	hw	-1.914	.805	5.655	1	.017	.147	.030	.714
	td	-2.860	1.093	6.845	1	.009	.057	.007	.488
	ew	1.508	.536	7.929	1	.005	4.518	1.582	12.909
	sl	.892	.310	8.261	1	.004	2.441	1.328	4.484
	end	-3.195	1.132	7.966	1	.005	.041	.004	.377
	nsd	2.009	.741	7.352	1	.007	7.459	1.745	31.880
	iod	.348	.834	.175	1	.676	1.417	.276	7.262
	ind	.314	.854	.136	1	.713	1.369	.257	7.297
	dsa	39.883	8.598	21.518	1	<.001	2.093E+17	10055256283.00	4.358E+24

a. The reference category is: rsph.

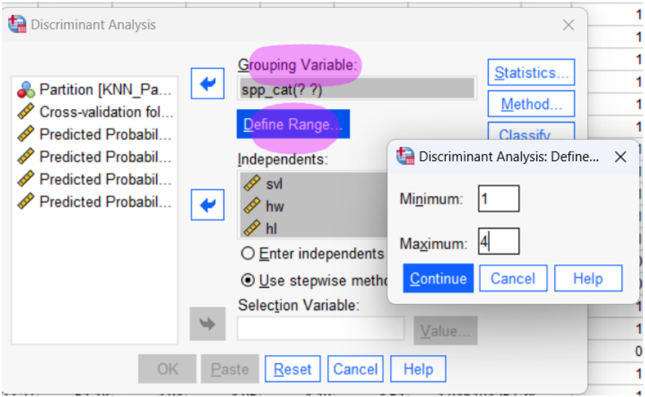
Lab 3

- Now let's run the LDA analysis. SPSS requires that the response be coded to numerical and check the Variable View to make sure that the Measure is `Scale`



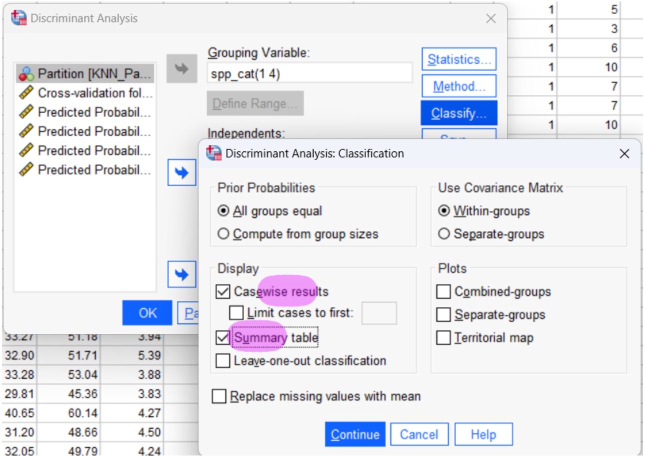
Lab 3

- click on `Analyze` → `Classify` → `Discriminant...`, and then add the Dependent variable and Covariates



Lab 3

- Click on `Classify` and set the desired classification options.



Lab 3

Standardized Canonical Discriminant Function Coefficients

	Function		
	1	2	3
svl	-.122	-.584	1.829
hl	-1.338	-.456	-1.311
td	-.867	-.362	.708
ew	.644	-.163	.644
sl	1.933	1.210	-1.189
end	.048	-.890	-.239
nsd	.450	.124	.344
iod	-.427	.065	.278
ind	-.079	.562	-.220
dsa	-.224	.639	-.320

Lab 3

Classification Results[a]

		spp_cat	Predicted Group Membership 1.00	2.00	3.00	4.00	Total
Original	Count	1.00	123	21	1	12	157
		2.00	5	20	5	0	30
		3.00	1	1	27	2	31
		4.00	4	5	0	37	46
	%	1.00	78.3	13.4	.6	7.6	100.0
		2.00	16.7	66.7	16.7	.0	100.0
		3.00	3.2	3.2	87.1	6.5	100.0
		4.00	8.7	10.9	.0	80.4	100.0

a. 78.4% of original grouped cases correctly classified.

License