Input Coding Techniques

- Simplicity and Generalisation
- Raw data
  - Geographic Information System example
- Data analysis
- Network design considerations
- Input coding decisions

Simplicity and Generalisation

- Simplest neural network which accounts for a data set will on average lead to the best generalisation to the population the data was drawn from.
  - too small network → over simplified solution
  - 'just right sized' → good generalisation
  - too large network → overfitting a danger, may memorise training cases
- Measures of complexity of networks
  - number of interconnections
  - number of processing neurons
  - number of bits per weights
- Weight decay
  - decrease weights during training – eliminates unnecessary connections
    - $w_{ij}(t + 1) = 0.9 * f$ where f is weight calculated normally for current model
- Network pruning
  - distinctiveness of hidden neurons

Coding input

- Generalisation is easiest if similar inputs have similar output vectors
  ⇒ avoid encodings which encrypt structure of the data
  ⇒ use encodings which enhance the structure
- Example: a real valued input variable

  i Binary number as 0/1 on a number of inputs
  - bad: least significant bits change often, but means little in most cases

  ii Single real value
  - better, but hard to treat different ranges differently

  iii Binary intervals: no. of binary inputs each represent a different range
  - lose discrimination within range

  iv Discrete intervals: no. of real inputs each encode a value in range
  - discontinuities at edges of ranges

  v Overlapping soft intervals
  - encourages generalisation to nearby values

Geographic Info System example

- Sample of raw data:

Pl	As	Sa	Ca	Al	Tp	Sl	Ge	Ra	Te	T1	T2	T3	T4	T5	T6	T7	Sc	Ds	Wd	Ws	Rf
1	0	50	99	36	96	50	90	39	60	55	20	20	48	30	36	26	10	90	10	10	10
2	60	35	35	31	64	70	90	39	60	56	18	16	41	20	33	18	10	90	10	10	10
3	50	50	0	21	32	50	90	39	60	56	21	21	61	44	48	32	10	90	10	10	10
4	50	50	0	35	32	70	90	39	60	57	20	17	63	34	39	24	10	10	10	90	10
5	10	50	99	43	96	50	90	39	60	58	19	19	46	38	36	30	10	90	10	10	10
6	80	35	85	33	48	40	40	39	60	61	21	24	66	40	36	36	10	90	10	10	10
7	80	35	85	34	96	40	40	39	60	57	19	18	57	25	36	20	10	10	10	90	10
8	70	0	50	30	32	40	40	39	60	57	19	18	51	33	36	22	10	10	10	90	10
9	80	35	85	36	64	60	40	39	60	62	22	24	49	42	36	40	10	90	10	10	10
10	80	35	85	32	32	50	40	39	60	60	20	21	48	33	36	28	10	10	10	90	10

- Have 190 points from rectangular grid of 244,494 points.

Geographic Info System example 2

- The raw data is a vector of 16 values: 7 taken from satellite images and augmented with information derived from a terrain model.
- The last 5 columns are the forest supra-type as determined from on ground observation.
  - This is time consuming, and expensive, which is why so few of the grid points are available labelled with the forest type.

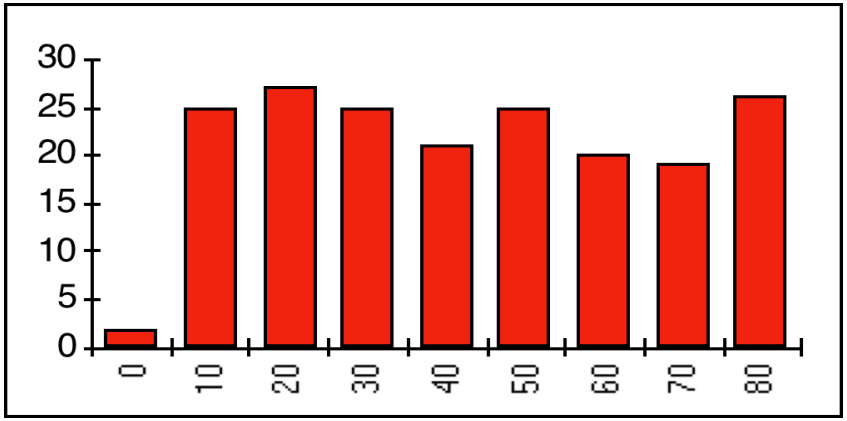
Geographic Info System example 3

- Plot number
- Aspect
- Sin of aspect
- Cos of aspect
- Altitude
- Topographic position
- Slope degree
- Geology descr.
- Rainfall
- Temperature
- Landsat band tm1
- Landsat band tm2
- Landsat band tm3
- Landsat band tm4
- Landsat band tm5
- Landsat band tm6
- Landsat band tm7
- Scrub
- Dry Sclerophyll
- Wet-dry sclero.
- Wet sclero.
- Rain forest

Available information:

- Aspect:
  - 0: flat, 10: North, 20: NE, 30: East, ..., 70: West, 80: NW.
- Topographic position:
  - 32: gully, 48: lower slope, 64: mid-slope, 80: upper slope, 96: ridge.
- Slope degree:
  - 10: < 1%, 20: < 2.15%, 30: < 4.64%, 40: < 10%, 50: < 21.5%, 60: < 46.4%, 70: < 100%, 80: > 100%
- Geology descriptor: unknown encoding
- Rainfall: (mm - 801)/5
- Temperature: (degrees - 11)*30
- Landsat tm bands 1 to 7: values in range 0 to 256

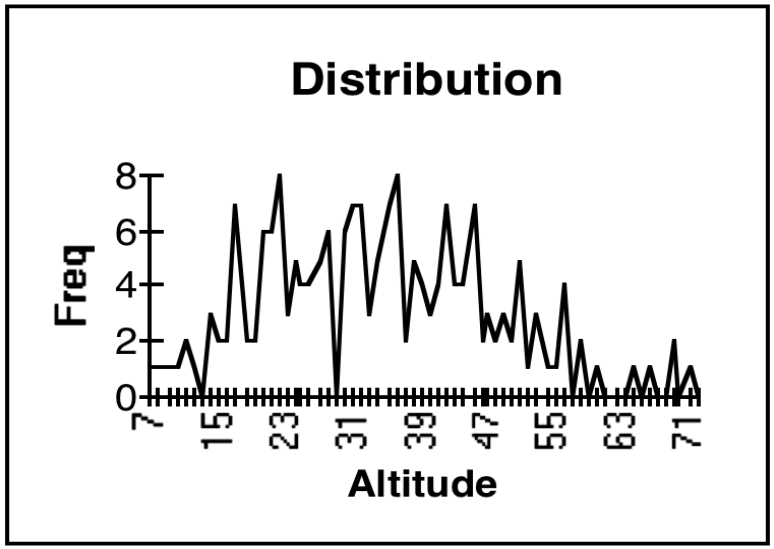
Data analysis – *Aspect, Sin & Cos*



- Distribution of *Aspect:*
  - These values encode the direction of the aspect of a particular geographic sample.
  - This is a ‘circular’ value, so AS=80, AS=10 may cause a net to generalise to 45!
  - Suited to either a continuous or category representation.
  - Choice to be made between the present encoding, another encoding based on sin and cos values only (these two functions can encode circular values) or category representation based on compass directions.
  - The 0 value for aspect is redundant since Slope Degree also encodes this info.

Data analysis – *Altitude*

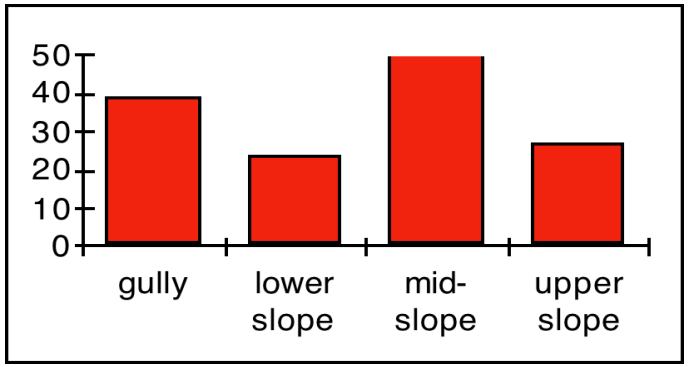
Maximum Value	71
Minimum Value	7
Average Value	34
Standard Deviation	13
Average Absolute Deviation	11



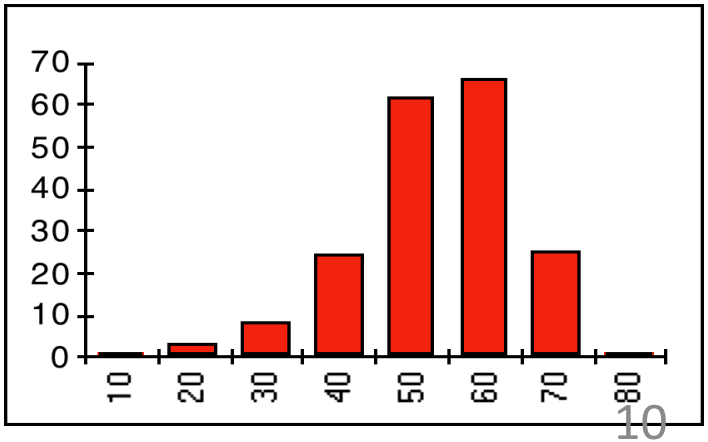
- Need to normalise altitude data over the range 0 - 1 for the network, since we are using the logistic function.
- Consider using statistical Z function to remove bias of lowest and highest values.

Data analysis – *Topog. & Slope*

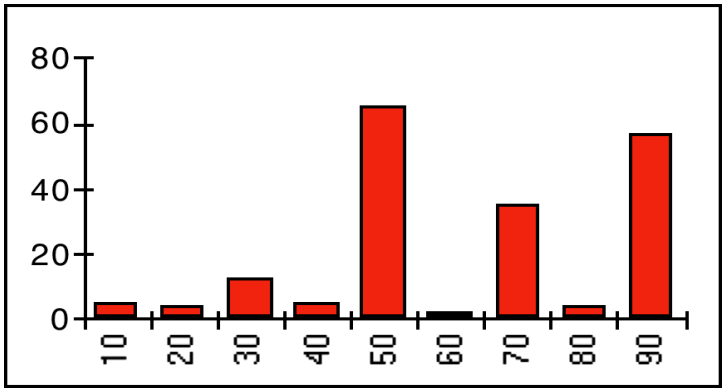
- Distrib. of *Topographic position:*
  - A category variable, appears to monotonically increase
    - it may be valuable to generalise between mid and upper slope for example.



  - Consider a continuous single input, or 3 or 4 category inputs.
  - Need to think about constructing the training and test sets with sensitivity to the different frequencies of occurrence in the data. E.g., mid-slope patterns are twice as frequent as lower slope.
- Distribution of *Slope degree:*
  - Consider using a single continuous valued input.



Data analysis – *Geology descriptor*



- Distribution of *Geology descriptor:*
  - There are really only 3 categories of major significance; the values of this indicator are not distributed normally.
  - Back-propagation trained neural networks usually perform better with linear or normal distributions to generalise from. Patterns which fall into the '50' band and are similar to patterns in the '70' band may cause spurious '60' results to be produced.
  - Consider using category inputs – each of the high frequency categories, and one extra for all the low frequency ones thrown together. Intuitively this means the input is a particular popular geography or indicated to be of unusual geography.
  - No information was available on the encoding used by the geographer on this field. Can not assume the variable is continuous given the distribution found above.

Network design considerations

- Topology
  - Need to keep the size of the network as small as possible to avoid overfitting the data. The more input, hidden or output neurons, the larger the number of weights and hence free parameters.
- Small size of training set – Noise
  - Consider adding random noise to training patterns to reduce chances of overfitting. Alternatively, consider adding (noisy) copies of existing patterns from categories with few training patterns.
- Back-propagation training parameters
  - Use ‘safe’ values initially.
  - Avoid momentum, use pattern mode (stochastic) updating of weights.

Network design considerations 2

- Output coding
  - Need to avoid problem of significantly different number of patterns in each class.
- Consider weighting patterns close to decision boundaries more heavily (if they can be identified).
- Small sample of the total data population – hence impossible to know if patterns in the training set and their frequency are representative of actual population. Task in constructing training and test sets is to provide sets that have similar distributions over the input vector.

Input coding decisions – *Aspect*

- The best coding for generalisation would be a category for each direction, with a redundant gaussian activating the input and the adjacent to a lesser amount. Would require 8 inputs (too many).
- The raw encoding of *aspect* does not reflect the circular nature of the information.
- As a compromise, code the *aspect* as 4 units, for each pt of the compass.
- Total activation for valid aspects is 2. The activation is apportioned in a way that should aid generalisation as there is a gradual change between directions.

As	Compass	A1	A2	A3	A4	Activ.
0	Flat	0	0	0	0	0
10	N	1	0.5	0	0.5	2
20	NE	1	1	0	0	2
30	E	0.5	1	0.5	0	2
40	SE	0	1	1	0	2
50	S	0	0.5	1	0.5	2
60	SW	0	0	1	1	2
70	W	0.5	0	0.5	1	2
80	NW	1	0	0	1	2

Input coding decisions – *misc.*

- *Altitude:*
  - Range from 7 to 71. Squash to range 0 - 1.
  - Distrib. of values is fairly normal, use a simple linear squashing function.
- *Topographic position:*
  - Raw indicator is relative and categorical. Ie. mid-slope is higher than gully and lower than upper slope.
  - Use a single continuous input:

gully	0.0	lower slope	0.25		
mid-slope	0.5	upper slope	0.75	ridge	1.0

- *Slope degree, Rainfall, Temperature:*
  - Use a single continuous input. Simple linear squash to range 0 - 1.
- *Landsat bands:*
  - Relating Landsat TM band values to forest supra-types is research.
  - Initially use a single continuous input, and simple linear squash to 0 - 1.
  - Expect to re-encode subsequently.

Input coding decisions – *Geology descriptor*

- There is no known relationship between the categories.
- From the data it appears to be a nominal value. There is no particular distribution, three of the types are quite common and the others are rare.
- A single continuously valued input is not appropriate. The categories will be represented by 4 inputs.
- The coding distinguishes between the popular types and collects the rare ones together, and losing some information. In any case, the input vector over these inputs would be quite sparse.
- All unmarked activations are set to 0.1. The cases that activate G1, are about equal in total to the other three categories.

Output Coding

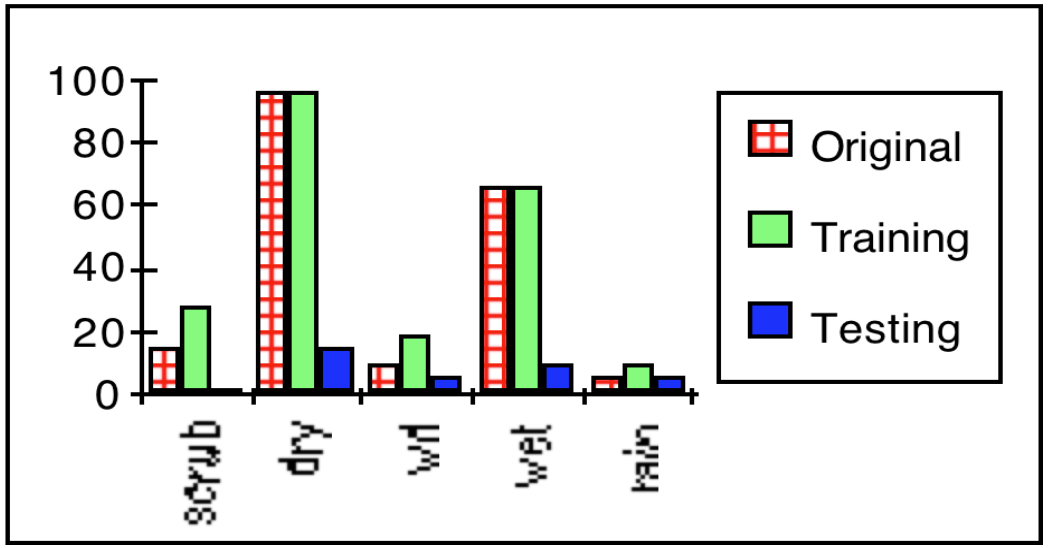
- Network is a classifier, and the output vectors are very sparse. Can lead to difficult learning, since the tss can improve by pushing output vectors to all 0's. Avoid by equilateral coding, so 5 possibilities represented by 4 units. All units have some activation on each pattern, maximum dist. between vectors maintained.
- To retrieve the category, calculate Euclidean distance between the output vector and the above values:

Category	Unit 1	Unit 2	Unit 3	Unit 4
Scrub	0.184	0.317	0.371	0.4
Dry Sclerophyll	0.816	0.317	0.371	0.4
Wet-dry sclero.	0.5	0.865	0.371	0.4
Wet sclerophyll	0.5	0.5	0.887	0.4
Rain Forest	0.5	0.5	0.5	0.9

$$\text{distance}(U_i) = \sqrt{\sum_{j=1}^{4} (y_j - u_{j_i})^2}$$

Training and test sets

- 190 available patterns
  - 160 patterns training set
  - 30 patterns test set
- Test patterns chosen at random, then modified to maintain proportions of each category of forest.
- Rare training patterns duplicated in training set.
- Distribution of the sets by category:



- The totals are: Original 190, Training 184, Testing 29.
  - Note how the profile for the testing set differs for the rare/common forest types.

Results

- 12 hidden unit network

epochs	tss - train	tss - test	actual perform. train/184	test/29
750	12.9	3.9	143	17
1050	11.4	4.4	151	16
1350	9.7	4.8	156	16
1650	9.1	5.0	156	15

- The network fails to learn any of the rainforest patterns. Note that generalisation suffers as more epochs are run.

Adding Random noise

- Trained with normal, and randomly distorted data in 50 epoch alternation. For example, at 200 epochs did 50 normal, then 50 @ 0.4 random distortion.

epoch	tss train	tss test	actual train	actual test	randomness
200	19.2	3.8	108	16	0.4
300	17.9	3.5	128	17	0.3
400	16.7	3.1	130	18	0.2
500	16.3	3.6	124	13	0.15
600	18.9	3.6	133	17	0.1
700	16.9	3.4	131	16	0.05
800	13.2	3.3	143	20	0.0

- Note crude form of simulated annealing used in the randomness profile. Note also tss vs. actual test result.