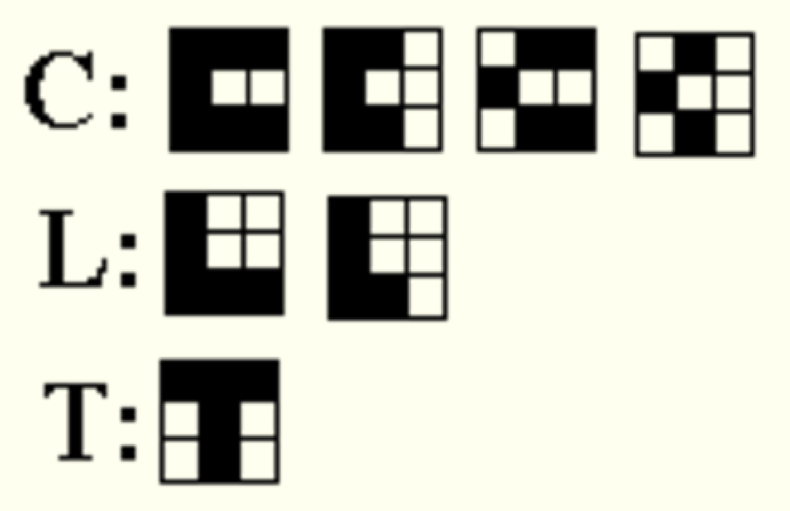
Tutorial 7

Friday 11th March, 2016

Problem 1. Design a multilayer perceptron which will learn to recognize various forms of the the letters C,L,T placed on a 3x3 grid through backpropagation algorithm.

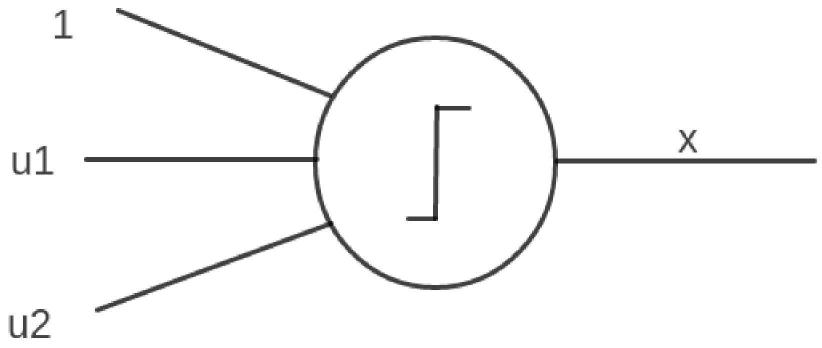


1. Design a one layer network indicating what should be applied at the input layer and what should be expected at the output layer showing the number of neurons, the connections between them and the neurons output function.

2. repeat (a) for two layer network by adding a hidden layer

Problem 2. Consider a perceptron for which $u \in R^2$ and

$$f(a) = \begin{cases} 1 & a > 0 \\ 0 & a = 0 \\ -1 & a < 0 \end{cases}$$



Let the desired output be 1 when elements of class A = {(1,2),(2,4),(3,3),(4,4)} is applied as input and let it be -1 for the class B = {(0,0),(2,3),(3,0),(4,2)}. Let the initial connection weights $w_0(0) = +1, w_1(0) = -2, w_2(0) = +1$ and learning rate be h = 0.5.

This perceptron is to be trained by perceptron convergence procedure, for which the weight update formula is $w(t+1) = w(t) + \eta(y^k - x^k(t))u^k$

1. (a) Mark the elements belonging to class A with x and those belonging to class B with o on input space.
   (b) Draw the line represented by the perceptron considering the initial connection weights w(0).
   (c) Find out the regions for which the perceptron output is +1 and 1
   (d) Which elements of A and B are correctly classified, which elements are misclassified and which are unclassified?
2. If u=(4,4) is applied at input, what will be w(1) ?
3. Repeat a) considering w(1).
4. If u=(4.2) is then applied at input, what will be w(2)?
5. Repeat 1) considering w(2).
6. Do you expect the perceptron convergence procedure to terminate? Why?

Problem 3. Let X and Y be *independent continuous* random variables with same density functions

$$p(x) = \begin{cases} e^{-x} & \text{if } x > 0; \\ 0 & \text{otherwise.} \end{cases}$$

Find density $\frac{X}{Y}$.

Problem 4. Use of Bayes' Theorem A lab test is 99% effective in detecting a disease when in fact it is present. However, the test also yields a false positive for 0.5% of the healthy patients tested. If 1% of the population has that disease, then what is the probability that a person has the disease given that his/her test is positive?

Problem 5. Part of Speech Tagging POS tagging is a problem of great importance in the field of Natural Language Processing, **NLP**. Refer to figure1.
Input: A set of n-words
Output: POS tag for each word

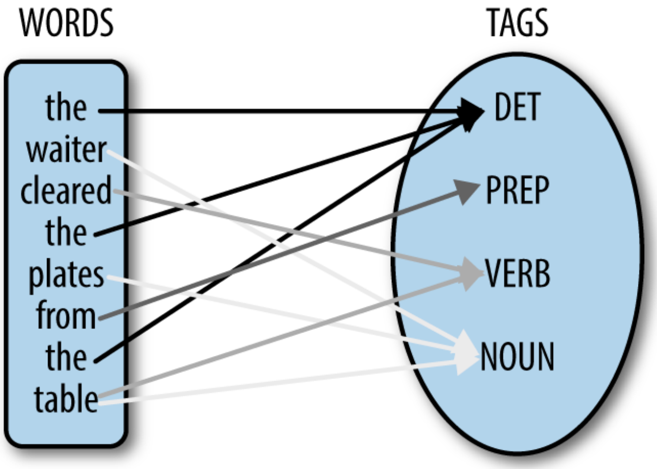


Figure 1: POS Tags

Assuming that the picking of words is done independently, find probability that the set contains a 'noun' given that it contains a 'verb'. Formulate the entire setting (sample space, events, your own probability distribution function or probability mass function, etc) to answer this question. .

Problem 6. For next lecture: Consider a binary classification problem where we are given a training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Let us assume that the positive examples are labelled as 1 and the negative examples are labelled as 0. Let us also model the probability that an example x belong to the positive class using sigmoid function as follows:

$$\theta = P(y = 1|x; w) = g(w^T\phi(x)) \text{ where } g(s) = \frac{1}{1+e^{-s}}. \tag{1}$$

Clearly, $P(y = 0|x; w) = 1 - g(w^T\phi(x)) = 1 - \theta$. It should be noted that w is the parameter of the model. In general, we can write $P(y|x; w)$ as

$$P(y|x; w) = \theta^y (1 - \theta)^{1-y}. \tag{2}$$

Then in logistic regression, the likelihood function of the parameter w is defined as

$$P(\mathcal{D}; w) = \Pi_{i=1}^m P(y_i|x_i; w) = \Pi_{i=1}^m \theta_i^{y_i} (1 - \theta_i)^{1-y_i} \tag{3}$$

where $\theta_i = P(y_i = 1|x_i; w)$. In maximum likelihood approach, one maximizes the likelihood function to estimate the value of the parameter w.

1. To maximize the likelihood $P(\mathcal{D}; w)$ with respect to w, one can minimize the negative log-likelihood $E(w) = -logP(\mathcal{D}; w)$ with respect to w. Derive the expression for $J(w)$.

2. $E(w)$ can be minimized with respect to w using gradient descend algorithm. Derive the expression of the gradient of $J(w)$ with respect to w.