SCS 3251 – Statistics for Data Science

Multiple and Logistic Regression

Course Roadmap

Module / Week	Title
1	Introduction to Statistics for Data Science
2	Probability
3	Distribution of Random Variables
4	Inference Part 1
5	Inference Part 2
6	Linear Regression
7	Multiple Regression
8	Logistic Regression
9	Introduction to Bayesian Inference
10	Multi-level Models
11	Markov Chain Monte Carlo
12	Presentations
13	Final Exam

Module 8: Learning Objectives

- Model evaluation
- Likelihood estimators
- Logistic Regression
  - Applications

Key Topic Overview – Multiple & Logistic Regression

- How to evaluate a model
- Principle of the Maximum Likelihood Estimator
- When to use Logistic Regression
- Logistic Regression introduction
- Logistic regression application to neural networks - introduction

MODEL EVALUATION (DISCRETE CHOICE)

Model Evaluation

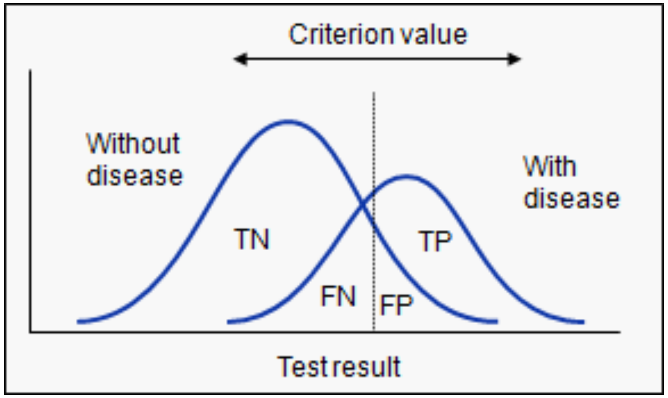
Confusion Matrix		Truth	
		TRUE	FALSE
Predicted	TRUE	Sensitivity (as a proportion)	Type I error False Alarm
	FALSE	Type II error Miss	Specificity (as a proportion)

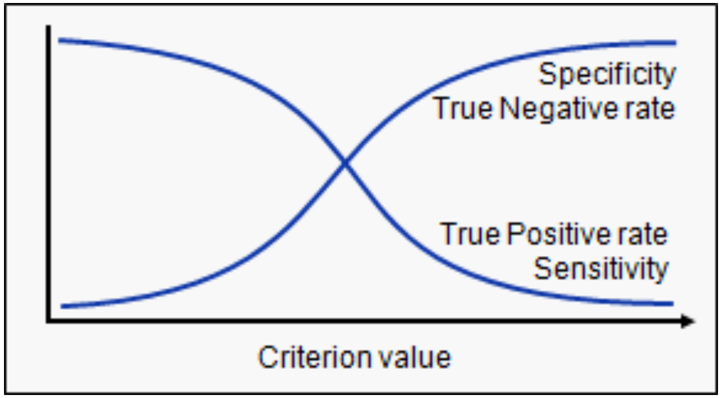
Confusion Matrix		Truth	
		TRUE	FALSE
Predicted	TRUE	True Positive TP	False Positive FP
	FALSE	False Negative FN	True Negative TN

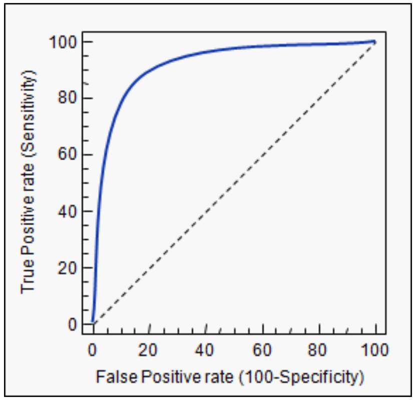
- The values in the confusion matrix will depend on the cut-off value
- *i.e. the probability above which we deem the output as a prediction of positive and below as negative*
- *There is a trade-off between sensitivity and specificity*

Model Evaluation

- A model is almost never able to predict the true positives and the true negatives with no error (misclassification)
- The cut off value is chosen with the help or the ROC curve (Receiver Operating Characteristic)
- The cut-off chosen will depend on the application and the risk appetite







https://www.medcalc.org/manual/roc-curves.php

Model Evaluation - Statistics

- Sensitivity : True Positive Rate = $\frac{TP}{TP+FN} = \frac{Positive\ Prediction}{Total\ Positives}$
- Specificity : True Negative Rate = $\frac{TN}{TN+FP} = \frac{Negative\ Prediction}{Total\ Negatives}$
- *Positive predictive value:* $\frac{TP}{TP+FP}$ *= probability that it is true when the prediction is true*
- *Negative predictive value =* $\frac{TN}{TN+FN} =$ *probability that it is false when the prediction is false*
- *Positive likelihood ratio =* $\frac{TPR}{FPR} = \frac{Sensitivity}{(1-Specificity)}$
- *Negative likelihood ratio =* $\frac{FNR}{TNR} = \frac{(1-Sensitivity)}{Speicificity}$

MAXIMUM LIKELIHOOD ESTIMATOR

Likelihood Principle

- It is not always clear what estimation method to use
- In such cases we use the likelihood guiding principle
- This principle leads to an estimation method called maximum likelihood, a standard method used to analyze data for many advanced statistical techniques
  - regression analysis, logistic regression analysis, time series analysis, categorical data analysis, survival analysis, and structural equation models[1]

1 Understanding Advanced Statistical Methods, Peter H. Westfall and Kevin S. S. Henning, CRC Press

Likelihood Function

- *The model we use when we use the Likelihood principle is:*
  - *model produces data*
  - that the *model has unknown parameter*
  - that *data reduce the uncertainty about the unknown parameters*.
- We want to find a model that is able to produce the data of our sample…

$$p(y|\theta) \rightarrow \mathrm{DATA}$$

- We assume our data to be independent and identically distributed (iid) then

$$p(y|\theta) \rightarrow Y_1, Y_2, \ldots, Y_n$$

$$p(y_1, y_2, \ldots, y_n|\theta) = p(y_1|\theta) \times p(y_2|\theta) \times \cdots \times p(y_n|\theta)$$

Likelihood Function – Cont.

- The likelihood function is defined as a function of ∅

$$L(\theta|y_1,y_2,\ldots,y_n) = p(y_1,y_2,\ldots,y_n|\theta), \quad \text{for } \theta \in \Theta$$

- Where ∅ represents all the possible values of θ, the parameter space

$$L(\theta|y_1,y_2,\ldots,y_n) = p(y_1|\theta)\times p(y_2|\theta)\times\cdots\times p(y_n|\theta)$$

- If $p(Y_1/\theta)$ follows a Bernoulli distribution

$$p(y\,|\,\pi) = \pi, \text{ if } y = 1, \text{ and } p(y\,|\,\pi) = 1-\pi,$$

- And for example we have data were 392 cases yield 1 and 610 cases yield 0

$$L(\pi|392 \text{ ones and } 610 \text{ zeros}) = \pi^{392}(1-\pi)^{610}$$

Maximum Likelihood Estimation

$X_1, X_2, X_3, \ldots, X_n$ have joint density denoted $f_\theta(x_1, x_2, \ldots x_n) = f(x_1, x_2, \ldots x_n | \theta)$

Given observed values $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, the likelihood of θ is the function

$$\text{likelihood}(\theta) = f(x_1, x_2, \ldots x_n | \theta)$$

Likelihood(θ) is the probability of observing the given data as a function of θ.

For independent and identically distributed data the likelihod simplifies to

$$\text{likelihood}(\theta) = \prod_i f(x_i | \theta)$$

Rather than maximizing this product, we often use the fact that the logarithm is an increasing function, so it is equivalent to maximize the log likelihood:

$$l(\theta) = \sum_i \log(f(x_i | \theta))$$

Maximum Likelihood Estimation

Poisson distribution: $P(X = x) = \lambda^x e^{-\lambda}/x!$

$$L(\lambda) = \sum_i (X_i \log\lambda - \lambda - \log X_i!) = \log\lambda \sum_i X_i - n\lambda - \sum_i \log X_i!$$

Take the first derivative to find maximum:

$$L'(\lambda) = 1/\lambda \sum_i x_i - n = 0,$$

Which implies that the estimate should be

$$\lambda = \langle X \rangle \text{ or sample average.}$$

Normal distribution: $\partial L/\partial\mu = 1/\sigma^2 \sum_i (x_i - \mu)$

$$\partial L/\partial\sigma = -n/\sigma + \sigma^{-3} \sum_i (x_i - \mu) 2$$

Limitations of MLE

- If the likelihood function has more than one peak, the numerical method might converge to the wrong peak, depending on the initial value.
- If the data and/or the model is inadequate, or if the likelihood function is very complicated, the method might not converge at all.
- If there are parameter constraints (e.g., variances must be positive), the usual methods can have trouble locating cases where the solution is on the boundary, where the derivative is not zero.

The Likelihood Ratio Test

- Likelihood provides an automatic, usually highly efficient method to estimate parameters.
- It is similarly useful for testing hypotheses: The *likelihood ratio test* provides tests that are also usually highly efficient
  - They have the greatest ability to detect deviations from the chance-only (or null) model
  - likelihood ratio tests are *optimal* in the sense of having the highest power among certain types of tests
  - *F*-statistic is a likelihood ratio statistic
  - likelihood ratio tests are useful because they give you a way to test hypotheses in *any* likelihood-based model, whether based on normal distributions, Poisson distributions, Bernoulli distributions

LOGISTIC REGRESSION

Why Logistic Regression?

- Simple and multiple linear regressions are applied when the response is continuous
- These models does not work well when the response is a categorical value with two levels
  - Is a customer going to buy an item (Y/N)?
  - Is an e-mail spam (Y/N)?
- These situations are better modeled by a type of generalized linear model (GLM): Logistic Regression
- GLMs can be thought of as a two-stage modeling approach
  - First model the response variable using a probability distribution, such as the binomial or Poisson distribution
  - Second, model the parameter of the distribution using a collection of predictors and a special form of multiple regression

Introduction to Logistic Regression

Logistic regression is a generalized linear model where the outcome is a two-level categorical variable. The outcome, Y_i, takes the value 1 when a condition is met with probability θ_i and the value 0 with probability $1 - \theta_i$ when the condition is not met.

- It is the probability p_i that we model in relation to the predictor variables.
- The logistic regression model relates the probability (p_i) to the predictors $x_{1,i}$, $x_{2,i}$, ..., $x_{k,i}$ through a framework much like that of multiple regression:

$$transformation(\theta_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$$

- Linear models require the residuals to be normally distributed, this is not possible in the case of a predicted binary variable. That is why (among other things) we use logistic models

Model Response Variable - Logit Transformation

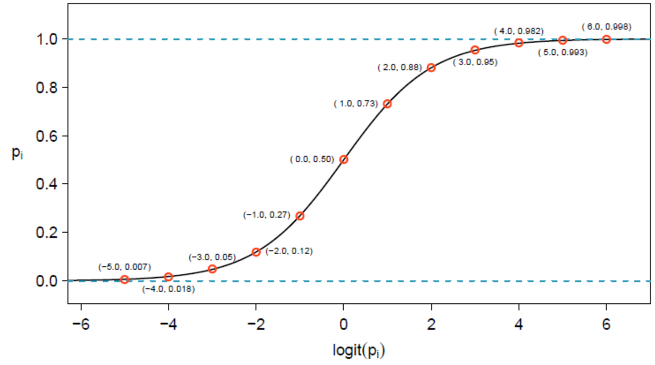
- The most common transformation for p_i is the logit transformation, which may be written as:

$$Logit(\theta_i) = Log_e(\frac{\theta_i}{1-\theta_i})$$

- $(\frac{\theta_i}{1-\theta_i})$ is what in colloquial English we know as the odds

- Solving for θ_i yields(θ_i is a sigmoid function):

$$\theta_i = \frac{e^{\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \cdots + \beta_k x_{k,i}}}{1 + e^{\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \cdots + \beta_k x_{k,i}}}$$



Model the response

- The common practice is to model the response variable taking values of 0 or 1 to represent a <u>no</u> or <u>yes</u> answer
- Predictors can be categorical or numerical variables
- If outliers are present in predictor variables, the corresponding observations may be especially influential on the resulting model
  - This is the motivation for omitting the numerical variables by transforming or binning the predictors

TIP: Notation for a logistic regression model
The outcome variable for a GLM is denoted by Y_i, where the index *i* is used to represent observation *i*.
The predictor variables are represented as follows: $x_{1,i}$ is the value of variable 1 for observation *i*, $x_{2,i}$ is the value of variable 2 for observation *i*, and so on.

Some Observations on Logistic Regression

- Point estimates will generally change a little - and sometimes a lot - depending on which other variables are included in the model. This is usually due to collinearity in the predictor variables.
- A positive coefficient estimate in logistic regression, just like in multiple regression, corresponds to a positive association between the predictor and response variables when accounting for the other variables in the model
- A positive coefficient indicates that the probability will increase with the presence of that characteristic

Logistic Model Interpretation

- Any classifier will have some error
- Every classifier will fall into one of three categories:
  1. The classifier output (based on input characteristics) indicates the absence of the attribute, typically when the output of the model is quite low, say, under 0.05.
  2. The characteristics generally indicate the presence of an attribute, the resulting probability is quite large, say, over 0.95.
  3. The characteristics roughly balance each other out in terms of evidence for and against the classifier and its probability falls in the remaining range
- Thresholds are of special importance as they impact the number of items with or without an attribute being correctly classified

Impact of Thresholds (Additional Material)

- Play with thresholds to see how the following matrix would be populated in the spam example
- What thresholds would you choose?

	Actual		
Predicted		True	False
	True	True Positives	False Positives
	False	False Negatives	True Negatives

- False Positives: good e-mails going to the spam folder
- False Negatives: spam e-mails going to the inbox

Diagnosis of a Logistic Regression

Logistic regression conditions

There are two key conditions for fitting a logistic regression model:

1. Each predictor x_i is linearly related to logit(p_i) if all other predictors are held constant.
2. Each outcome Y_i is independent of the other outcomes.

- As we said before, the output is a categorical variable so we cannot impose the normal distribution of the error as a condition or requirement

Modelling a Binomial Event

Let's first recap the binomial process:

1. There are *m* identical trials
2. Each trial results in one of two outcomes, either a "success," *S* or a "failure," *F*
3. θ , the probability of "success" is the same for all trials
4. Trials are independent

Let *Y* = number of successes in *m* trials of a binomial process. Then *Y* is said to have a binomial distribution with parameters *m* and *q* .

$Y \sim \mathrm{Bin}(m, q)$ with the properties: $E(Y) = m\,\theta$, $Var(Y) = m\,\theta(1-\theta)$

$$P(Y=j) = \binom{m}{j}\theta^{j}(1-\theta)^{m-j} = \frac{m!}{j!(m-j)!}\theta^{j}(1-\theta)^{m-j} \qquad j = 1,\ldots,m$$

Modelling a Binomial Event (Cont'd)

- In the logistic regression we wish to model the proportion of successes on the basis of predictors $x_1, x_2, \ldots, x_n$

$$(Y|x_1, x_2, \ldots, x_n) \sim \text{Bin}(m_i, \theta(x_{1i}, x_{2i}, \ldots, x_{ni}))$$

$$y_i / m_i = \theta(x_{1i}, x_{2i}, \ldots, x_{ni})$$

- With the following conditions
  - y_i / mi is an unbiased estimate of $\theta(x_{1i}, x_{2i}, \ldots, x_{ni})$
  - y_i / mi varies between 0 and 1

Please note that the expected value and the variance of the response depend on θ, as such they are not constant.

$Y \sim \text{Bin}(m, q)$ with the properties: $E(Y) = m\,\theta$, $Var(Y) = m\,\theta(1 - \theta)$

Thus, least squares regression is an inappropriate technique for analyzing Binomial responses

Binomial Event Example

- We will compare to restaurant guides, Michelin Guide New York City and Zagat Survey 2006: New York City Restaurants
- We want to be able to model θ , the probability that a French restaurant is included in the 2006 Michelin Guide New York City , based on customer views from the Zagat Survey 2006: New York City Restaurants
- Data:
  - m_i restaurants in the Zagat Survey received
  - x_i food rating, of those m_i,
  - y_i are listed in the Michelin guide (successes),
  - m_i - y_i *are not in the Michelin guide (failure)*

Food rating, x_i	InMichelin, y_i	NotInMichelin, m_i-y_i	m_i	y_i/m_i
15	0	1	1	0.00

Binomial Event Example (Cont'd)

- Use the 'MichelinFood.txt' and 'Module7 LogisticRegressionBinomialExample.ipynb' files
- From the example, the fitted model is

$$\hat{\theta}_i = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x)}} = \frac{1}{1+e^{-(-10.842+0.501x)}}$$

- Rearranging, the log of the odds or logit is

$$Logit(\theta_i) = Log_e\left(\frac{\theta_i}{1-\theta_i}\right) = -10.842 + 0.501x$$

- In logistic regression the concept of the residual sum of squares is replaced by a concept known as the deviance defined to be

$$G^2 = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{y}_i}\right)\right]$$

where

$$\hat{y}_i = m_i\hat{\theta}$$

$$\text{df} = n - (\text{number of } \beta\text{'s estimated})$$

Binomial Event Example (Cont'd)

- From the software/example:

Null deviance: 61.427 on 13 degrees of freedom

Residual deviance: 11.368 on 12 degrees of freedom

Chi Square test of deviance vs. residuals = 0.498

$P(G2 > 11.368) = 0.498$

- we fail to reject the Null hypothesis, that the model is appropriate
- Note: different models are compared by comparing the deviances using a chi square distribution

Binomial Event Example (Cont'd)

- Recall that for linear regression

$$R^2 = 1 - \frac{\text{RSS}}{\text{SST}}.$$

Since the deviance can be written as

where S are the successes in the data and M are from the model

$$R^2_{\text{dev}} = 1 - \frac{G^2_{H_A}}{G^2_{H_0}} \qquad G^2 = 2\left[\log(L_S) - \log(L_M)\right]$$

In this specific example

$$R^2_{\text{dev}} = 1 - \frac{11.368}{61.427} = 0.815$$

The difference in these two deviances and its p-value is given by

$$G^2_{H_0} - G^2_{H_A} = 61.427 - 11.368 = 50.059 \qquad P(G^2_{H_0} - G^2_{H_A} > 50.059) = 1.49\text{e-}12$$

Modelling a Binary Event

- A binary event can be thought as a special binomial event where m_i = 1
- in this situation the goodness-of-fit measures X^2 and G^2 are not good measures to evaluate/compare models
- When modelling a single event the response is either 0 or 1, i.e. absence or presence of an attribute, as a function of the predictor/s
- We will work an example with one predictor
- Notice that when m=1, the log-likelihood function isgiven by

$$\log(L)=\sum_{i=1}^{n}\left[y_i\log\left(\theta(x_i)\right)+\left(1-y_i\right)\log\left(1-\theta(x_i)\right)+\log\binom{1}{y_i}\right]$$

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Data Normalization

- In order to avoid bias in the response the independent variables should all have the same or similar range
- This is critical if the data ranges are very different, like when measured on different scales
- Some techniques used are:
  – z score of the data
  – Proportion of the min-max range (%)
  – Sigmoidal or Softmax, apply the sigmoid curve to the z value of the data
  – Bining
  – Principal components

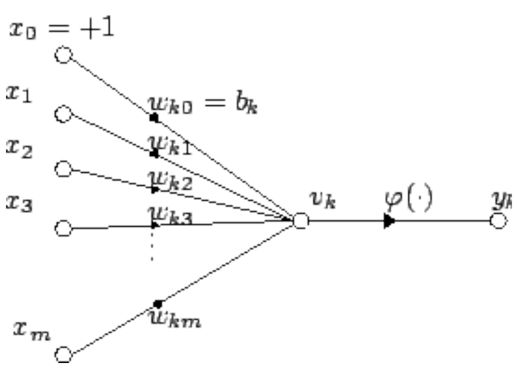
INTRODUCTION TO ARTIFICIAL NEURAL NETWORKS

Artificial Neural Network

- Artificial neural networks, called neural nets for short, model the unknown function by expressing it as a weighted sum of several sigmoids, usually chosen to be logit curves, each of which is a function of all the relevant explanatory variables.

Artificial Neuron

- A simple model of the neuron is a switch that based on inputs from other neurons decides if to fire or not
- The output of the kth neuron is modelled as

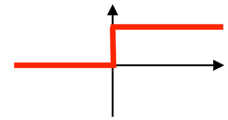
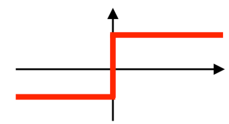
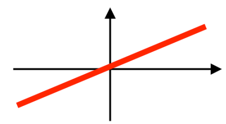
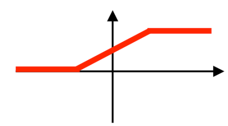
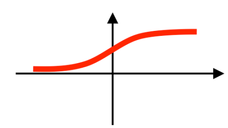
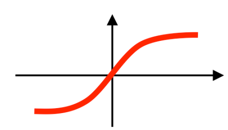


$$y_k = \varphi\left(\sum_{j=0}^{m} w_{kj} x_j\right)$$

- The function φ is called the transfer function, the most common one is the sigmoid $output_k = \frac{1}{1+e^{-(\sum w_{ik}+w_{0k})}}$
- However there are more

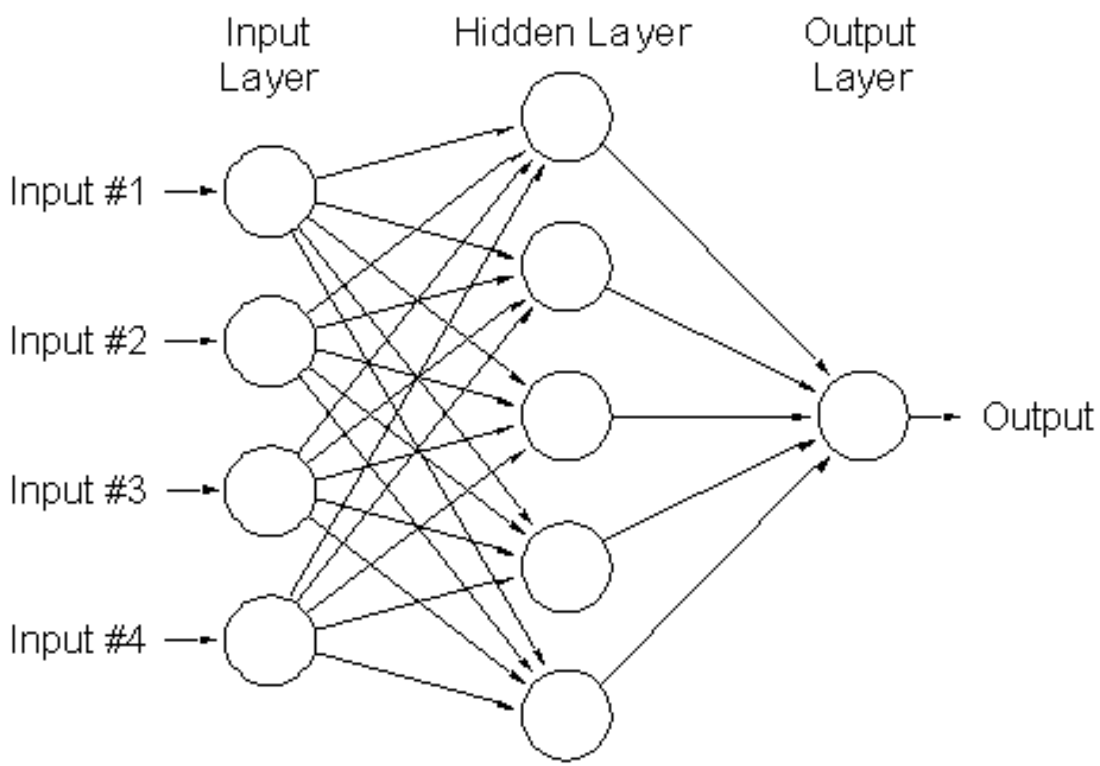
https://en.wikipedia.org/wiki/Artificial_neuron

Transfer Function

Activation function	Equation	Example	1D Graph
Unit step (Heaviside)	$\phi(z) = \begin{cases} 0, & z < 0, \\ 0.5, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Sign (Signum)	$\phi(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Linear	$\phi(z) = z$	Adaline, linear regression	
Piece-wise linear	$\phi(z) = \begin{cases} 1, & z \geq \frac{1}{2}, \\ z + \frac{1}{2}, & -\frac{1}{2} < z < \frac{1}{2}, \\ 0, & z \leq -\frac{1}{2}, \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	Logistic regression, Multi-layer NN	
Hyperbolic tangent	$\phi(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$	Multi-layer NN	

https://www.quora.com/What-is-the-role-of-the-activation-function-in-a-neural-network#!n=18

The Neural Network



https://www.cs.bgu.ac.il/~ben-shahar/Teaching/Computational-Vision/StudentProjects/ICBV061/ICBV-2006-1-TorIvry-ShaharMichal/index.php

Before we go…

“All models are wrong, but some are useful”-George E.P. Box

The truth is that no model is perfect. However, even imperfect models can be useful. Reporting a awed model can be reasonable so long as we are clear and report the model's shortcomings.

Caution: Don't report results when assumptions are grossly violated

While there is a little leeway in model assumptions, don't go too far. If model assumptions are very clearly violated, consider a new model, even if it means learning more statistical methods or hiring someone who can help.

Further Reading

Both logisticmetric examples were taken from:
S.J. Sheather, A Modern Approach to Regression with R,
DOI: 10.1007/978-0-387-09608-7_1, © Springer Science + Business Media LLC 2009

This book offers a very good coverage of different types of regression, including topics not covered here such as transformation of variables. At some points it may be to 'mathy' for this course

MLE taken from Understanding Advanced Statistical Methods, Peter H. Westfall Kevin S. S. Henning, CHAPMAN & HALL/CRC (chapter 12 and 17 were used to build this deck)

This book is between a basic and an advanced course. As in intermediate course it is very good to clarify concepts

Next Class

- Bayes: from the reference

https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers

- Read chapter 1

https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/blob/master/Chapter1_Introduction/Ch1_Introduction_PyMC3.ipynb