DR ADRIAN BEVAN

PRACTICAL MACHINE LEARNING

REGRESSION PROBLEMS WITH NEURAL NETWORKS

LECTURE PLAN

- Regression
- Examples
  - Linear regression
  - MNIST
  - *B* physics background suppression
- Summary

QMUL Summer School: https://www.qmul.ac.uk/summer-school/
Practical Machine Learning QMplus Page: https://qmplus.qmul.ac.uk/course/view.php?id=10006

REGRESSION

- Using a model for regression means that we compute a quantitative output, called the *response*.
  - Essentially the process is the same as addressing a classification problem, however instead of a quantitative outcome (a yes/no decision on types associated with an example) we have a number associated with a data example.
  - This number (the response) can be used in a probabilistic sense to avoid having to make an absolute choice between types.
- All of the aspects of the problem encountered so far can be adopted for a regression analysis of a given problem.

REGRESSION

- Neural network models can be thought of as non-linear regression functions; a generalisation of a linear regression model.

- The examples we have encountered so far are processing abstract data samples such as images and scientific data; but we can use regression models as function approximators.

- To explain what this means we can consider a Taylor series expansion of some function.

REGRESSION

- e.g. consider the function y=f(x) = sin(x)
- We can model this with a polynomial; but which one?
  - $y=a_1x$
  - $y=a_1x+a_2x^2$
  - $y=a_1x+a_2x^2+a_3x^3$
  - $y=a_1x+a_2x^2+a_3x^3+a_4x^4$
  - $$y = \sum_{i=0}^{\infty} a_i x^i$$

Each model is an approximation of the function we are interested in.

For some region of x the approximation will be good (accurate prediction).

For some region in x the approximation will be bad (inaccurate prediction).

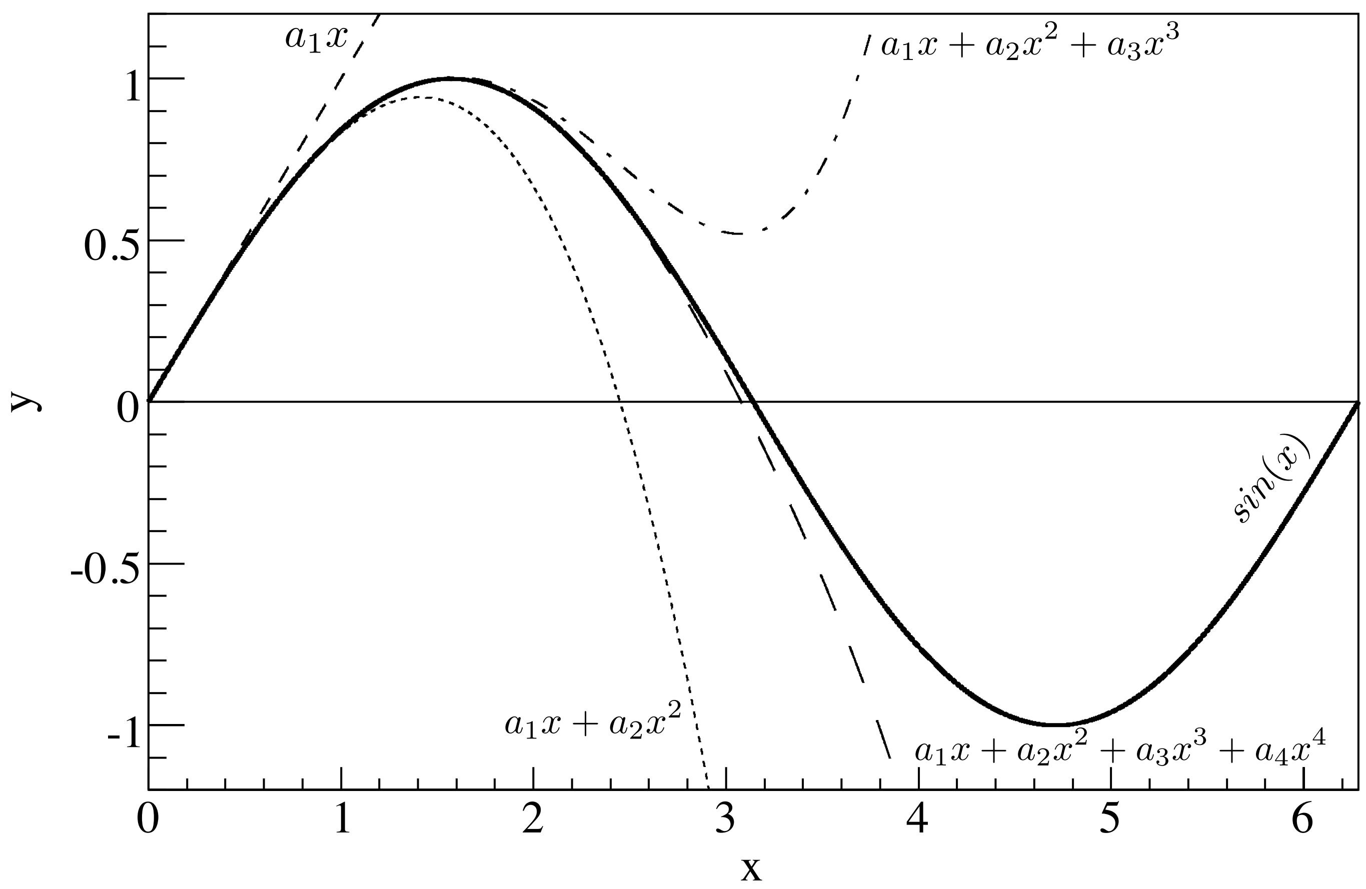
Like all approximations, there are conditions associated with the validity of a given model

We can neglect the a_0 term for this function approximation problem as sin(x) is an odd function (i.e. y=0 zero at x=0).

REGRESSION

- e.g. consider the function y=f(x) = sin(x)
- We can model this with a polynomial; but which one?
  - $y=a_1x$
  - $y=a_1x+a_2x^2$
  - $y=a_1x+a_2x^2+a_3x^3$
  - $y=a_1x+a_2x^2+a_3x^3+a_4x^4$
  - $$y = \sum_{i=0}^{\infty} a_i x^i$$

Range of validity increases with complexity of the model

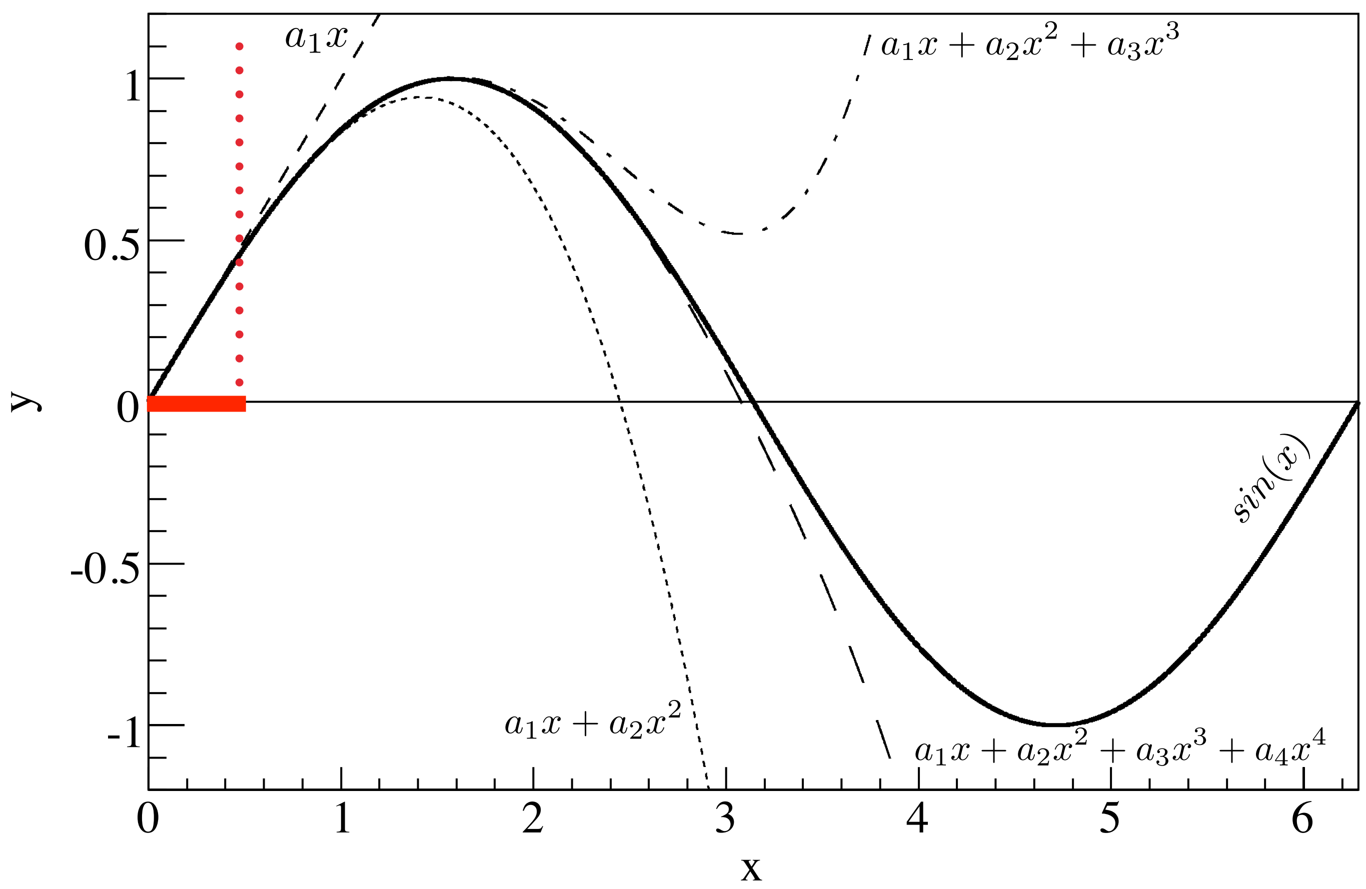


We can neglect the a_0 term for this function approximation problem as sin(x) is an odd function (i.e. y=0 zero at x=0).

REGRESSION

- e.g. consider the function y=f(x) = sin(x)
- We can model this with a polynomial; but which one?
  - $y=a_1x$
  - $y=a_1x+a_2x^2$
  - $y=a_1x+a_2x^2+a_3x^3$
  - $y=a_1x+a_2x^2+a_3x^3+a_4x^4$
  - $$y=\sum_{i=0}^{\infty}a_ix^i$$

Range of validity increases with complexity of the model →

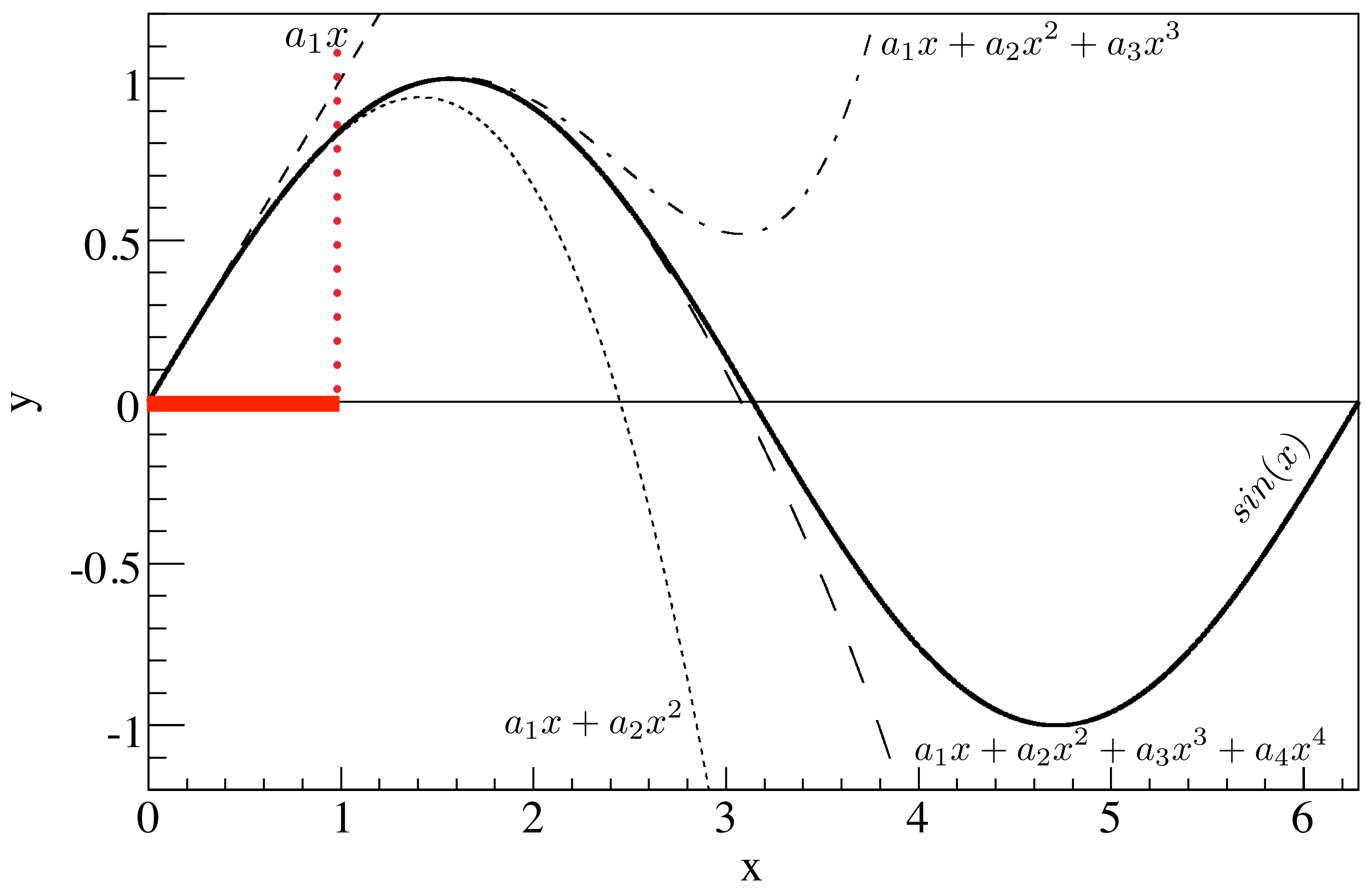


We can neglect the a_0 term for this function approximation problem as sin(x) is an odd function (i.e. y=0 zero at x=0).

REGRESSION

- e.g. consider the function y=f(x) = sin(x)
- We can model this with a polynomial; but which one?
  - $y=a_1x$
  - $y=a_1x+a_2x^2$
  - $y=a_1x+a_2x^2+a_3x^3$
  - $y=a_1x+a_2x^2+a_3x^3+a_4x^4$
  - $$y = \sum_{i=0}^{\infty} a_i x^i$$

Range of validity increases with complexity of the model

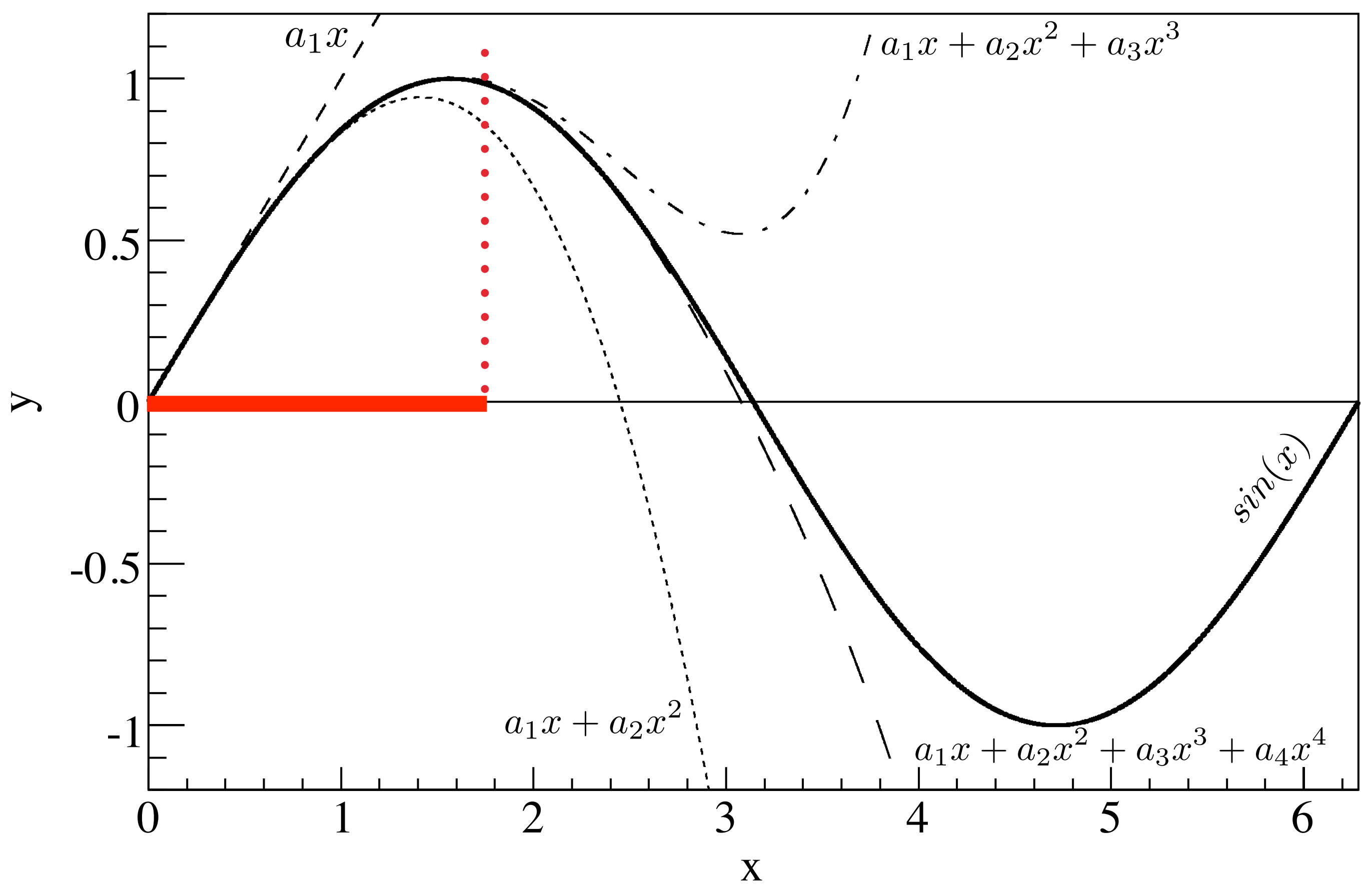


We can neglect the a_0 term for this function approximation problem as sin(x) is an odd function (i.e. y=0 zero at x=0).

REGRESSION

- e.g. consider the function y=f(x) = sin(x)
- We can model this with a polynomial; but which one?
  - $y=a_1x$
  - $y=a_1x+a_2x^2$
  - $y=a_1x+a_2x^2+a_3x^3$
  - $y=a_1x+a_2x^2+a_3x^3+a_4x^4$
  - $y = \sum_{i=0}^{\infty} a_i x^i$

Range of validity increases with complexity of the model

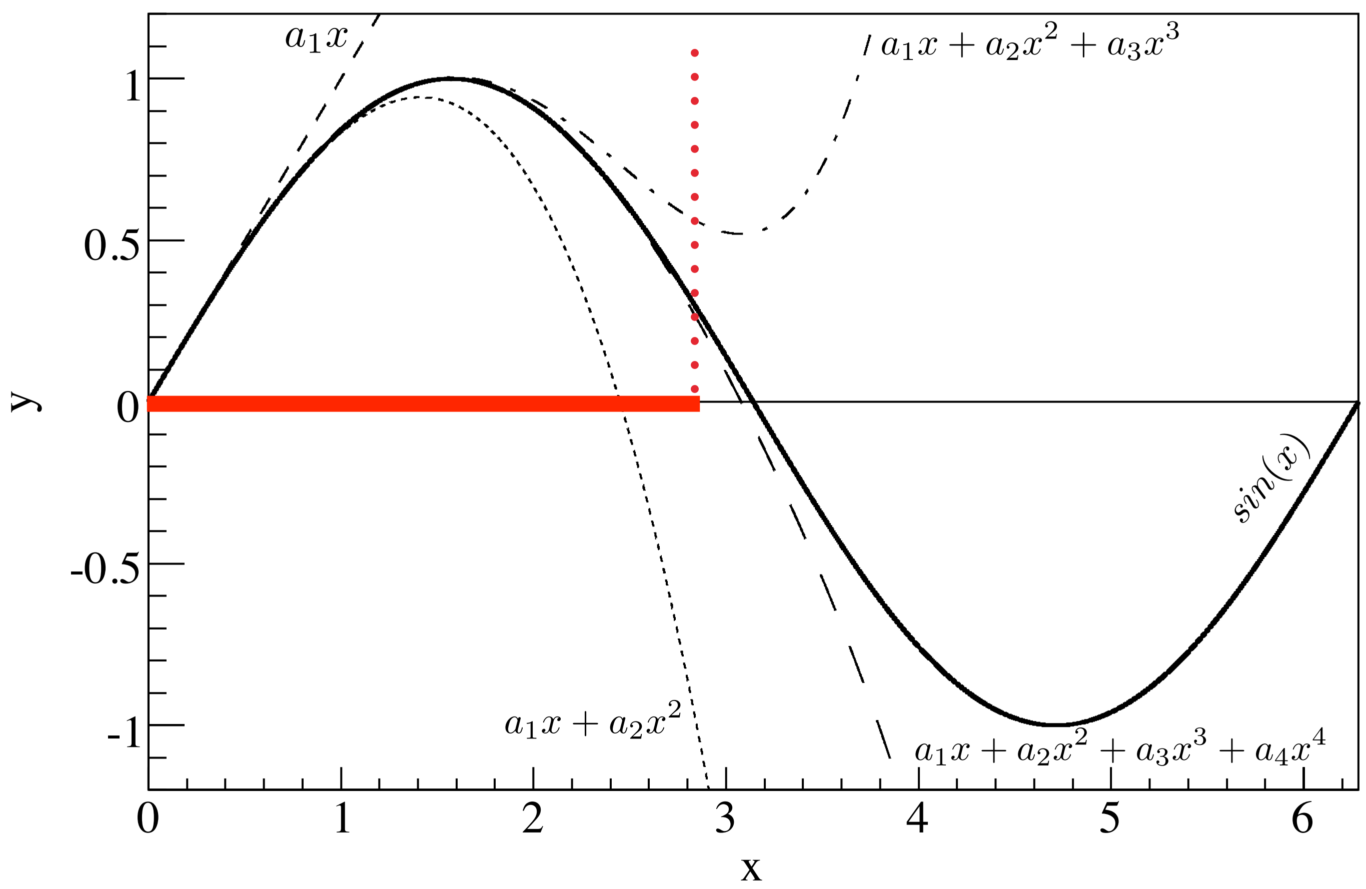


We can neglect the a_0 term for this function approximation problem as sin(x) is an odd function (i.e. y=0 zero at x=0).

REGRESSION

- e.g. consider the function $y=f(x) = \sin(x)$
- We can model this with a polynomial; but which one?
  - $y=a_1x$
  - $y=a_1x+a_2x^2$
  - $y=a_1x+a_2x^2+a_3x^3$
  - $y=a_1x+a_2x^2+a_3x^3+a_4x^4$
  - $$y = \sum_{i=0}^{\infty} a_i x^i$$

Range of validity increases with complexity of the model



We can neglect the a_0 term for this function approximation problem as sin(x) is an odd function (i.e. y=0 zero at x=0).

REGRESSION

- e.g. consider the function y=f(x) = sin(x)
- We can model this with a polynomial; but which one?

Taylor series expansions are analytically determined.

The analogy that a more complicated model can provide a better approximation to a function can down with a supervised learning approach:

If there are not enough training examples to determine the model parameters, then a complicated neural network can provide a bad approximation of a function.

We will come back to this when talking about optimisation, (training and overfitting).

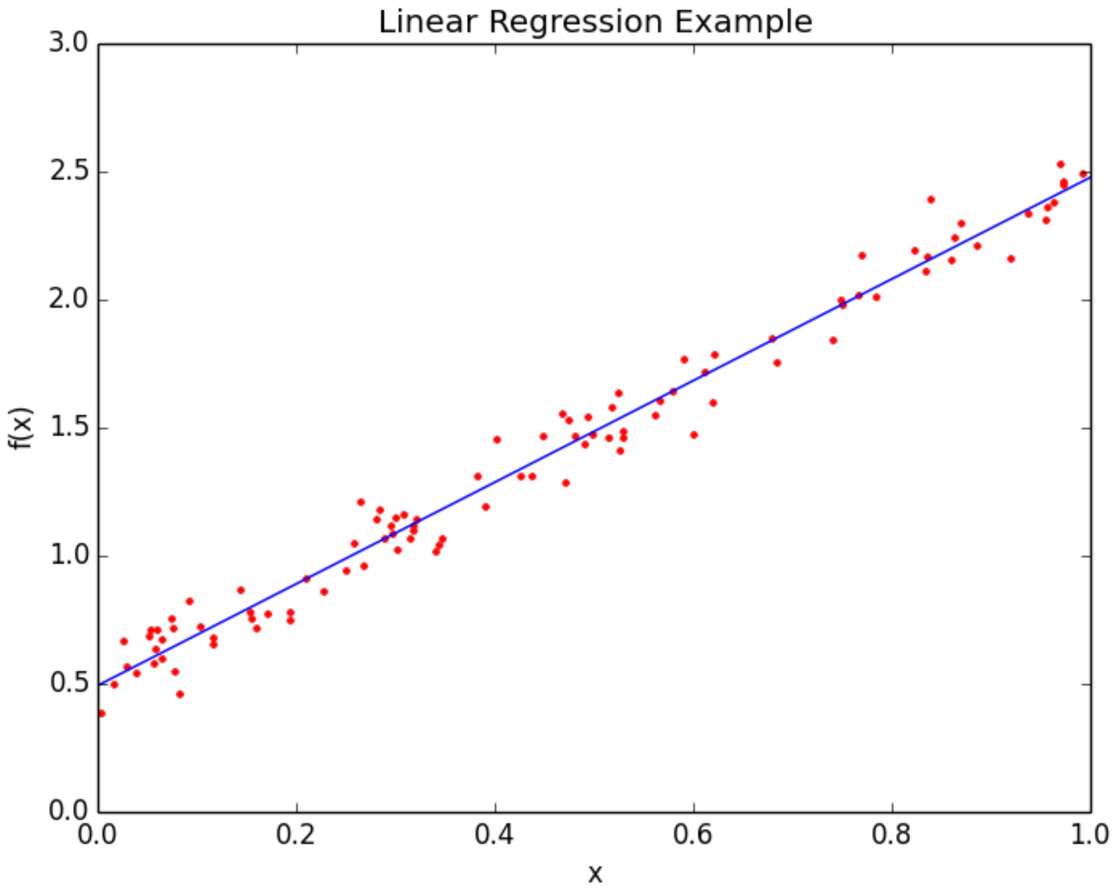
We can neglect the a_0 term for this function approximation problem as sin(x) is an odd function (i.e. y=0 zero at x=0).

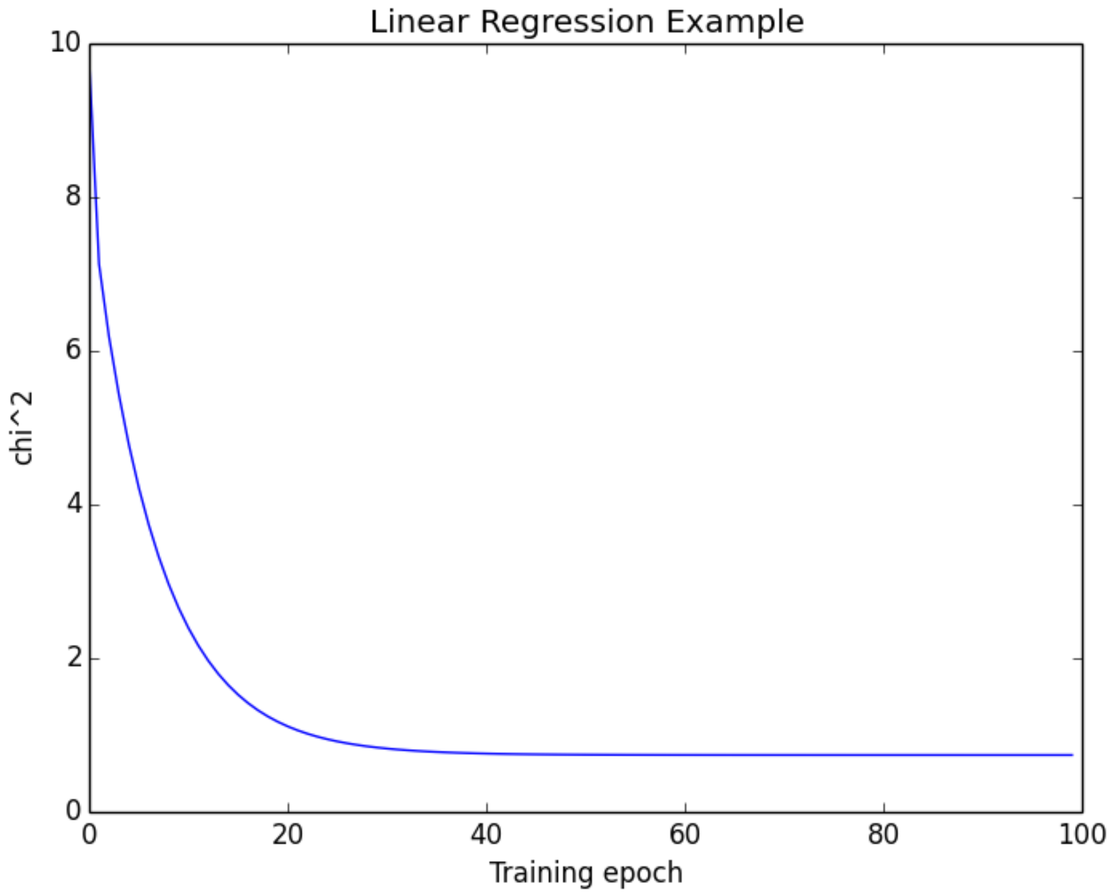
EXAMPLES: LINEAR REGRESSION

- We have already seen the linear regression problem; using optimisation to determine the model parameters for

$$y = f(x)$$

$$mx + c$$

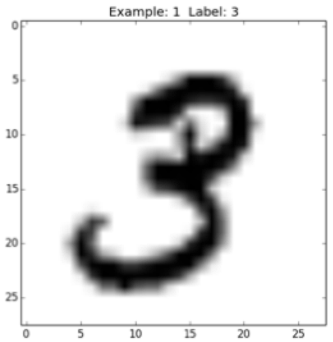


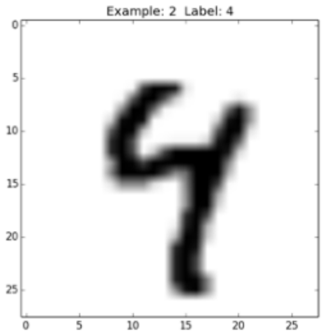


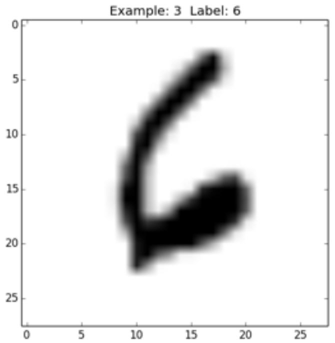
The model y is a prediction of a continuous output; this is a regression model prediction.

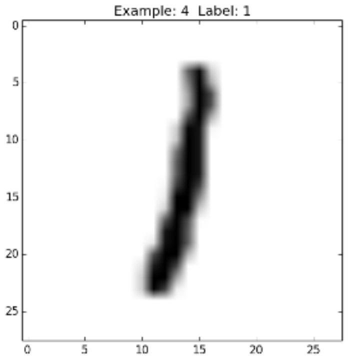
EXAMPLES: MNIST

- Consider the MNIST examples we looked at previously.









- A classification problem will require assigning a type to each example.
  - The error rates for various models summarised by LeCun were discussed in the previous lecture.
  - What if a model has two almost equiprobable assignments of type?
    - For such examples you may decide that it is not reasonable to make an absolute choice using a classification algorithm.
    - Instead you could consider assigning a relative probability for a given outcome (likelihood) for a particular assignment in order to preserve information about other possible assignments for the example.

[1] Neural Computation, Volume 22, Number 12, December 2010
http://yann.lecun.com/exdb/mnist/

EXAMPLES: B PHYSICS BACKGROUND SUPPRESSION

- Two experiments called *B Factories** were built to search for matter-antimatter asymmetry in *B* meson decays
  - A *B* meson is a $q\overline{q}$ pair where one quark is a *b* quark and the other is either a *u* or a *d* (for the purpose of this discussion).
  - Sometimes the matter and antimatter particles can decay at different rates. When this happens the symmetry CP is said to be violated (C=charge conjugation and P is parity). This is vital for understanding the Universe.
  - There was a model that naturally incorporated this phenomenon and could be used to predict CP violation in B decays.
- The work of these experiments validated the model of Kobayashi and Maskawa, that extended a earlier concept of quark mixing introduced by Cabibbo. The 2008 Nobel Prize included an award to Kobayashi and Maskawa for explaining this phenomenon.

*BaBar at the SLAC National Accelerator Laboratory in California, and Belle at the KEK Laboratory in Tsukuba, Japan. See, for example Bevan et al, arXiv:1406.6311.

EXAMPLES: B PHYSICS BACKGROUND SUPPRESSION

- The *B* factories collide electrons and positrons at a specific energy to make two *B* mesons nearly at rest in the centre of mass frame.

- The energies of the electrons and positrons are asymmetric to enable the study of a certain type of CP violation.

- Lets consider one particular example[1]: $B_d^0 \to \rho^+\rho^-$

- where $\rho^\pm \to \pi^\pm\pi^0$.

- Dominant background comes from $e^+e^- \to (u\overline{u}, d\overline{d}, s\overline{s})$.

- With preselection we have 1 signal event to O(70) background.

[1] Aubert et al., Phys.Rev.D76:052007,2007

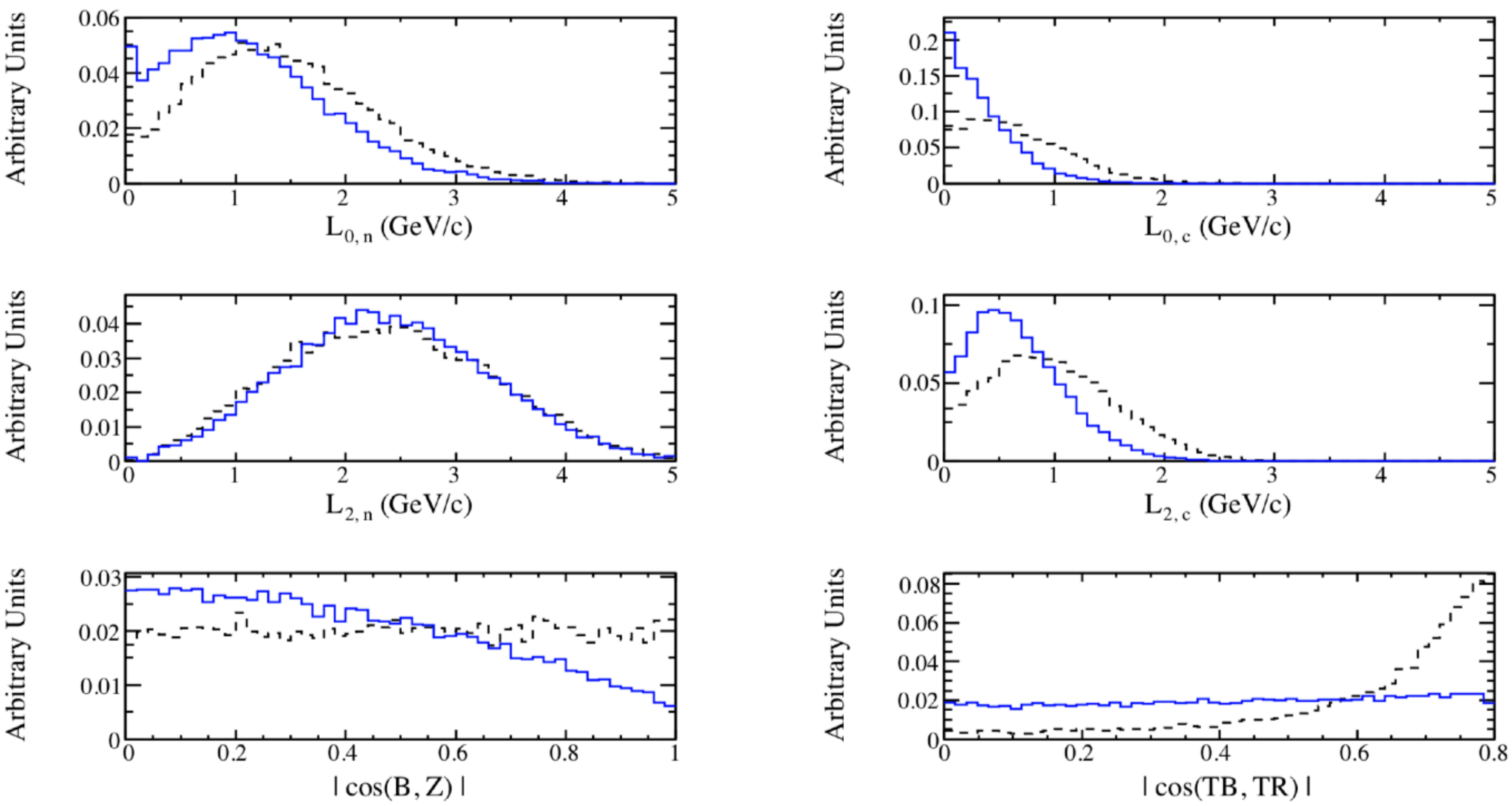
EXAMPLES: B PHYSICS BACKGROUND SUPPRESSION

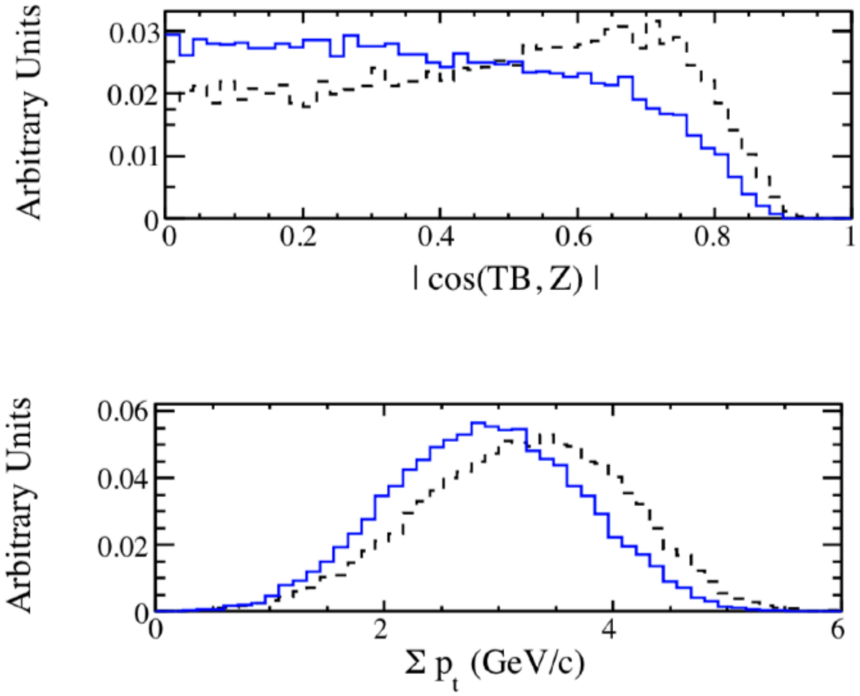
- Use an MLP to develop a regression model to compute a score for an example (event) to be signal or background.
- Distributions of known signal/background samples are used to construct a probability density function that is used in a likelihood fit to extract more information.
- Configuration of the network:
  - 8 dimensional input features;
  - 8 input nodes: two hidden layers of 7 and 6 nodes respectively, and one output node;
  - Use Sigmoid activation functions.

[1] Aubert et al., Phys.Rev.D76:052007,2007

EXAMPLES: B PHYSICS BACKGROUND SUPPRESSION

- Input features are based on energy flow of the event - the B mesons decay homogeneously whereas the light quark pair decays in a more focussed grouping.





The exact definitions of these variables are not important as you can understand these as all ways of looking at energy flow in a decay, or how spherically symmetric a decay is. See [1, 2] for more details.

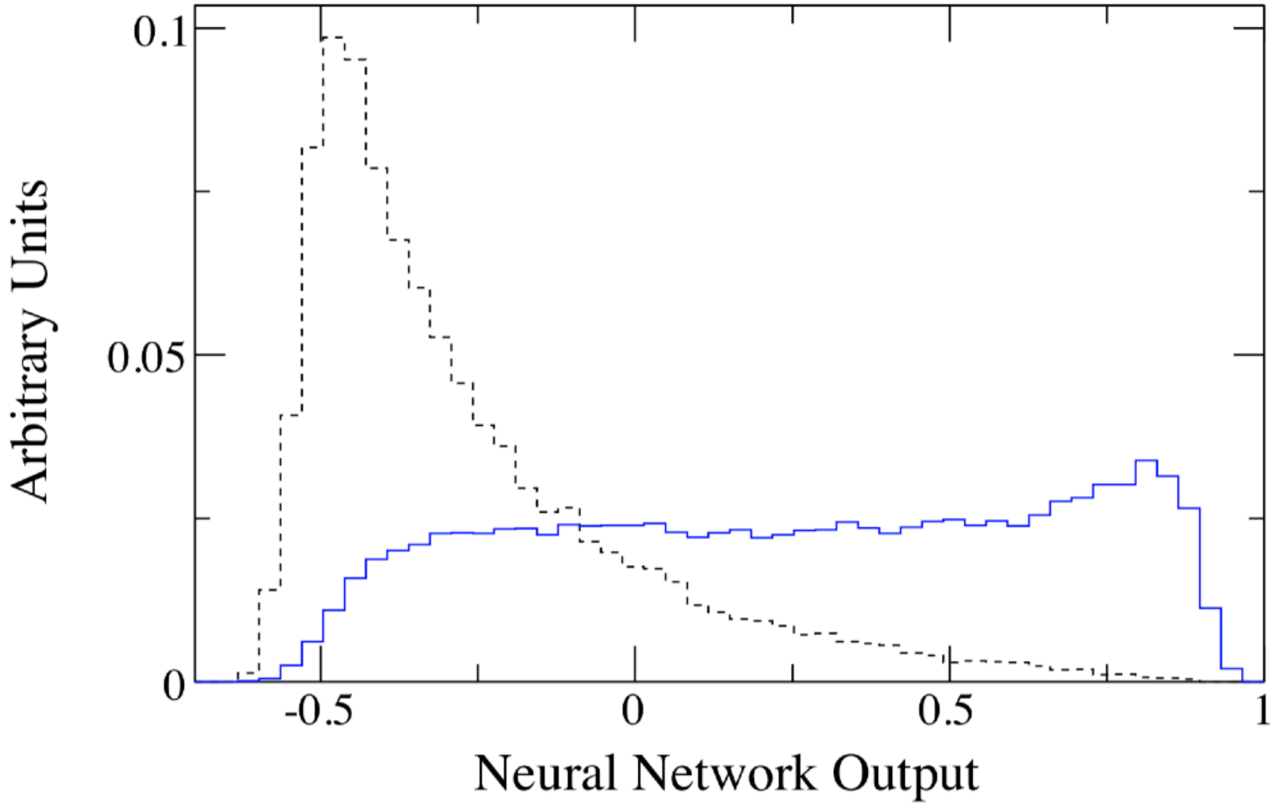
[1] Aubert et al., Phys.Rev.D76:052007,2007

[2] Ch 9 of Bevan et al, arXiv:1406.6311

EXAMPLES: B PHYSICS BACKGROUND SUPPRESSION

- The output of the MLP is transformed using a 1:1 mapping: $\mathcal{N} = 1 - \arccos(x - \xi); \;\; \xi = 0.0027$

- The reason for this is to spread out the signal distribution so that it can be modelled for the likelihood fit (otherwise it would just be a sharp peak near 1).



The mapping just spreads out the data, it does not add or destroy information provided by the model.

This is required because of the way that the output is used to obtain the main results of the paper.

[1] Aubert et al., Phys.Rev.D76:052007,2007

SUMMARY

- Regression analysis of data outputs a quantitative score for each example.
  - Can be used to construct a probability that an event is of one type or another.
  - Can be used to compare predictions and understand if the model significantly prefers one type over another, or if separation is marginal.
- Regression output of a model can be used in subsequent analysis.
  - Background suppression example used an MLP to provide data to construct probability density functions for further analysis.

SUGGESTED READING

- Discussion of regression models in text books
  - C. Bishop: *Neural Networks for Pattern Recognition*
    - Chapter: 1, 2, 6
  - C. Bishop: *Pattern Recognition and Machine Learning*
    - Chapter: 1, 2
  - T. Hastie, R. Tibshirani, J. Friedman, *Elements of statistical learning*
    - Chapter: 2, 3, 6
- A few examples of regression used in particle physics:
  - B Physics using an MLP: Aubert et al., Phys.Rev.D76:052007,2007
  - Exotic Particle Search using deep networks: Baldi, Sadowski, Whiteson, Nature Comm. 5308
  - Jet substructure: Baldi et al. Phys.Rev. D93 (2016) no.9, 094034
  - … many other examples out there to contextualise regression problems.