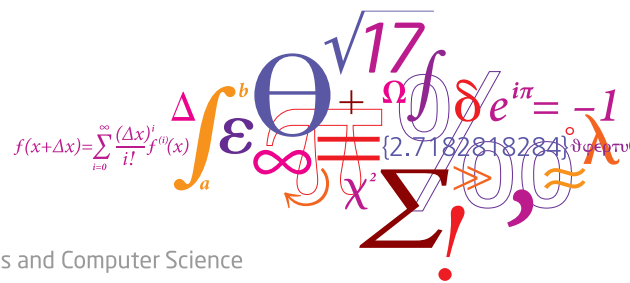
02450: Introduction to Machine Learning and Data Mining

Artificial Neural Networks and Bias/Variance

Jes Frellsen

DTU Compute, Technical University of Denmark (DTU)



DTU Compute
Department of Applied Mathematics and Computer Science

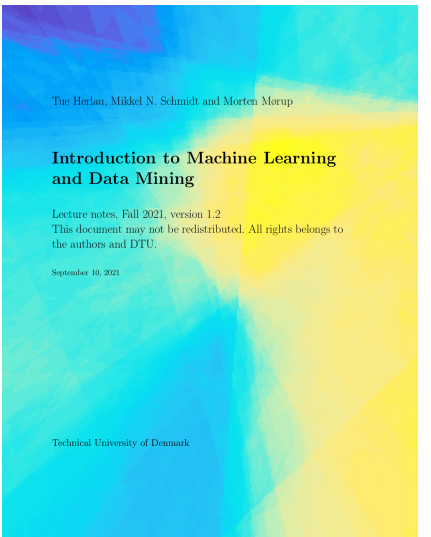
Today

Feedback Groups of the day:

Alexander Lambrecht, Andrea Arduin, Andreas Fabian Emiliussen Nøbbe, Aske Folkmann Musaeus, August Brogaard Tollerup, Aurelio Mottes, Carlo Antonello, Carlo Meroni, Changrun Liu, Charidimos Vradis, David Matthew Lane, Emil Block Overgaard, Felix Høgstædt Larsen, Frederik Engelsborg von Voss, Jakub Piatek Rzeznik, Joakim Bøegh Levorsen, Jonas Søeborg Nielsen, Julien Lacour, Junxuan Shi, Kristian van Kints, Lena Schlüter Nielsen, Line Sandvad Nielsen, Linnéa Haugen, Lorena Torres Lahoz, Mads Dyrved Møller, Magnus Leander Ovason, Mei Lin Verghese Law Kung Sam, Melina Siskou, Mikkel Aaby Kruse, Mikkel Amtrup Krohn, Mikkel Koefoed Lindtner, Mikkel Pedersen, Miriam Emilie Hart, Márton Ferenc Leitold, Nichlas Olesen, Nikolaos Stefanidis, Oliwia Lucyna Lasak, Patrik Kucerka, Peter Meyer Nielsen, Rebekka Steinhart, Roneet Vijay Nagale, Saxo Kilde Jessen Spiele, Siff Kasane Heike Ravn, Steen Nørkjær Larsen, Stefan Skrydstrup Pedersen, Stefanos Rodopoulos, Thor Gabriel Krøgholt-Damasceno, Victor Anton Charles Leweke, Youyang Shen, Yvet Maathuis

Reading material:

Chapter 14, Chapter 15



Lecture Schedule

1. Introduction
   31 August: C1

Data: Feature extraction, and visualization

2. Data, feature extraction and PCA
   7 September: C2, C3
3. Measures of similarity, summary statistics and probabilities
   14 September: C4, C5
4. Probability densities and data visualization
   21 September: C6, C7

Supervised learning: Classification and regression

5. Decision trees and linear regression
   28 September: C8, C9
6. Overfitting, cross-validation and Nearest Neighbor
   5 October: C10, C12 **(Project 1 due before 13:00)**
7. Performance evaluation, Bayes, and Naive Bayes
   12 October: C11, C13
8. **Artificial Neural Networks and Bias/Variance**
   26 October: C14, C15
9. AUC and ensemble methods
   2 November: C16, C17

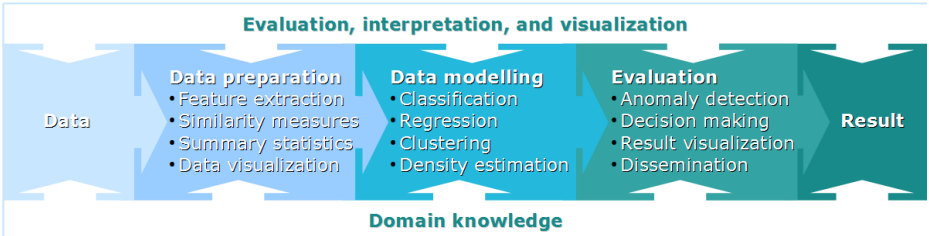
Unsupervised learning: Clustering and density estimation

10. K-means and hierarchical clustering
    9 November: C18
11. Mixture models and density estimation
    16 November: C19, C20 **(Project 2 due before 13:00)**
12. Association mining
    23 November: C21

Recap

13. Recap and discussion of the exam
    30 November: C1-C21

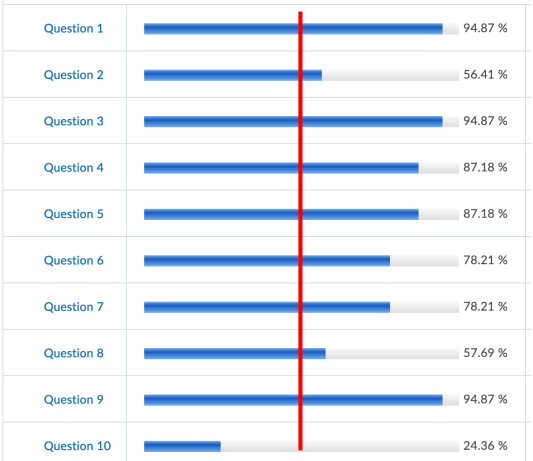
Online help: Forum on DTU Learn
Videos of lectures: `https://video.dtu.dk`
Streaming of lectures: Zoom (link on DTU Learn)

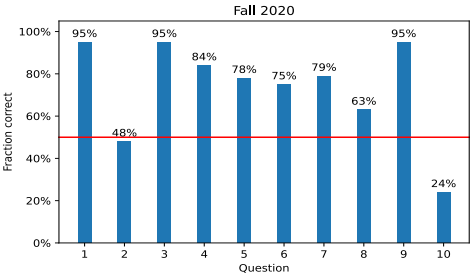


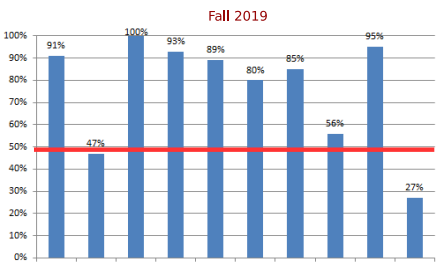
Learning Objectives

- Understand the Bias-Variance decomposition
- Understand and apply regularized least squares regression (i.e. ridge regression)
- Understand the principles behind artificial neural networks (ANNs) and how ANNs can be used for classification and regression
- Understand how logistic regression and ANNs can be extended to multi-class classification

Midterm practice test results







Solutions are at the end of this presentation

Question 2:
Consider the classification problem given in figure 1 and the Decision Tree in figure 2 with two decisions denoted A and B. We will let $\boldsymbol{x}_n$ define the x_1 and x_2 coordinates of a given observation whereas $\boldsymbol{x}_n$ -0.5·**1** denotes the subtraction of 0.5 from x_1 and x_2.
Which one of the following classification rules would lead to a correct classification of the data?

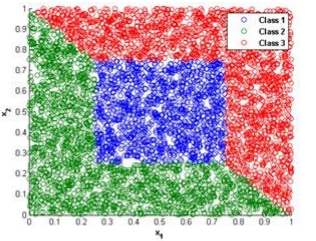


Figure 1

A: A: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_1 \leq 0.25$, B: $\|\boldsymbol{x}_n\|_\infty \leq 1$

B: A: $\|\boldsymbol{x}_n\|_1 \leq 1$, B: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq \infty$

C: A: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq 0.25$, B: $\|\boldsymbol{x}_n\|_\infty \leq 1$

D: A: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_\infty \leq 0.25$, B: $\|\boldsymbol{x}_n\|_1 \leq 1$

E: Don't know

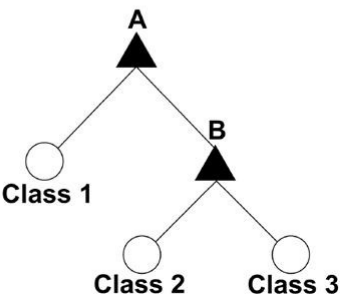
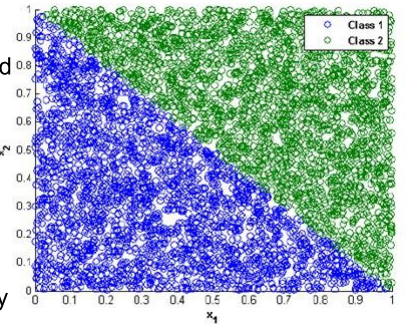


Figure 2

Question 8:
When carrying out a principal component analysis of a dataset with four attributes we obtain the following singular values s_1=4, s_2=2, s_3=1, and s_4=0. Which one of the following statements is wrong?

A: The first principal component accounts for more than 60 % of the variation in the data.
B: The third principal component accounts for less than 5 % of the variation in the data.
C: The second principal component accounts for more than 20 % of the variation in the data.
D: The data can be perfectly represented in a three dimensional sub-space.
E: Don't know.

Question 10:
Consider the classification problem given in Figure 5 where x_1 and x_2 are used as features for a logistic regression classifier and a decision tree. The considered logistic regression models all include the constant term w_0. Which one of the following statements is wrong?



A: The two classes can be perfectly separated by a logistic regression model using x_1 and x_2 as features.

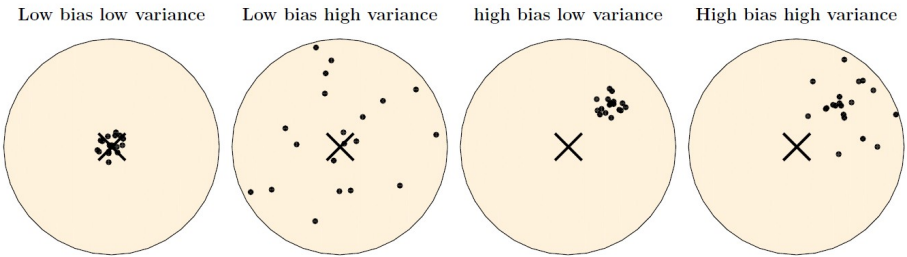
B: A decision tree with less than five nodes can perfectly separate the classes using only x_1 and x_2 as features.

C: A logistic regression model can perfectly separate the two classes using only the feature t given by $t = x_1 + x_2$.

D: In logistic regression the probability that each observation belong to the two classes can be derived from the logistic function.

E: Don't know.

What is bias and what is variance?



Regularized least squares

- Recall cost function from linear regression

$$E(\boldsymbol{w}) = \left\|\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{w}\right\|^2$$

- A parsimonious model can be obtained by **forcing** parameters towards zero.
- Problem: Columns of $\boldsymbol{X}$ have very different scale (i.e. require large/small values of $\boldsymbol{w}$)
- Therefore, standardize $\boldsymbol{X}$:

$$\hat{X}_{ij} = \frac{X_{ij} - \mu_j}{\hat{s}_j}, \quad \mu_j = \frac{1}{N}\sum_{i=1}^{N} X_{kj}, \quad \hat{s}_j = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_{ij} - \mu_j)^2}$$

- Note $\hat{\boldsymbol{X}}$ contains no constant term.

- Introduce regularization term $\lambda\|\boldsymbol{w}\|^2$ to penalize large weights:

$$E_\lambda(\boldsymbol{w}, w_0) = \sum_{i=1}^{N}(y_i - w_0 - \hat{\boldsymbol{x}}^\top \boldsymbol{w})^2 + \lambda\|\boldsymbol{w}\|^2 = \left\|\boldsymbol{y} - w_0\mathbf{1} - \hat{\boldsymbol{X}}\boldsymbol{w}\right\|^2 + \lambda\|\boldsymbol{w}\|^2$$

- We can solve for w_0 and $\boldsymbol{w}$:

$$\frac{dE_\lambda}{dw_0} = \sum_{i=1}^{N} -2(y_i - w_0 - \hat{\boldsymbol{x}_i}^\top \boldsymbol{w}) = -2N\mathbb{E}[y] - 2Nw_0 - N\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{x}_i}^\top\right)\boldsymbol{w}$$

$$\Rightarrow w_0 = \mathbb{E}[y]$$

- With $\hat{y}_i = y_i - \mathbb{E}[y]$

$$E_\lambda = \left\|\hat{\boldsymbol{y}} - \hat{\boldsymbol{X}}\boldsymbol{w}\right\|^2 + \lambda\|\boldsymbol{w}\|^2$$

- Setting the derivative wrt. $\boldsymbol{w}$ equal to zero and solving for $\boldsymbol{w}$ yields

$$\boldsymbol{w}^* = (\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}} + \lambda\boldsymbol{I})\backslash(\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{y}})$$

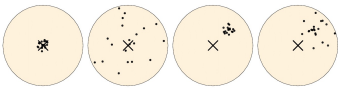
Selecting λ

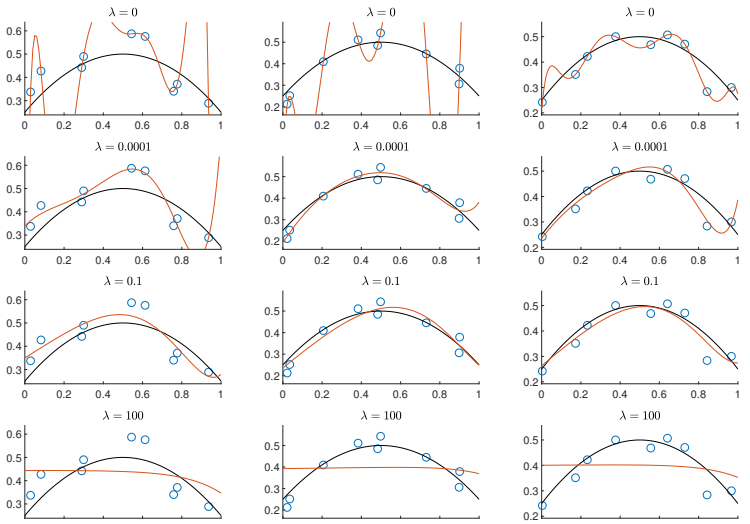
- Suppose

$$\boldsymbol{w}^* = (\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{X}} + \lambda \boldsymbol{I}) \backslash (\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{y}}) \propto \frac{Xy}{X^2 + \lambda}$$

- So if $\lambda = 0$ then no effect, else if $\lambda \to \infty$ then $\boldsymbol{w}^* \to 0$
- λ controls complexity of model. Select λ using cross-validation

How does different values of λ (vertical) affect the bias/variance of learned function (red lines)





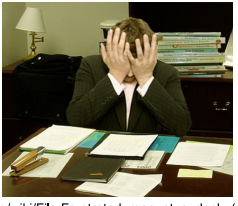
The Bias-Variance decomposition

$$\mathbb{E}_{\mathcal{D}}\left[E^{\text{gen}}\right] = \mathbb{E}_{\mathcal{D},(\boldsymbol{x},y)}\left[\left(y - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right]$$

We first consider $\boldsymbol{x}$ fixed

$$\bar{y}(\boldsymbol{x}) = \mathbb{E}_{y|\boldsymbol{x}}\left[y\right]$$

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D},y|\boldsymbol{x}}\left[\left(y - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D},y|\boldsymbol{x}}\left[\left(y - \bar{y}(\boldsymbol{x}) + \bar{y}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] \\
&= \mathbb{E}_{y|\boldsymbol{x}}\left[\left(y - \bar{y}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_{\mathcal{D}}\left[\left(\bar{y}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] + 2\mathbb{E}_{\mathcal{D},y|\boldsymbol{x}}\left[\left(y - \bar{y}(\boldsymbol{x})\right)\left(\bar{y}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x})\right)\right]
\end{aligned}$$



https://commons.wikimedia.org/wiki/File:Frustrated_man_at_a_desk_(cropped).jpg

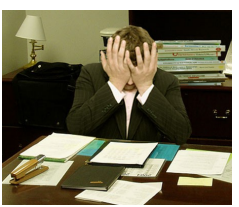
The Bias-Variance decomposition

$$\mathbb{E}_{\mathcal{D}}\left[E^{\text{gen}}\right] = \mathbb{E}_{\mathcal{D},(\boldsymbol{x},y)}\left[(y - f_{\mathcal{D}}(\boldsymbol{x}))^2\right]$$

We first consider $\boldsymbol{x}$ fixed

$$\bar{y}(\boldsymbol{x}) = \mathbb{E}_{y|\boldsymbol{x}}\left[y\right]$$

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D},y|\boldsymbol{x}}\left[(y - f_{\mathcal{D}}(\boldsymbol{x}))^2\right] \\
&= \mathbb{E}_{\mathcal{D},y|\boldsymbol{x}}\left[(y - \bar{y}(\boldsymbol{x}) + \bar{y}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x}))^2\right] \\
&= \mathbb{E}_{y|\boldsymbol{x}}\left[(y - \bar{y}(\boldsymbol{x}))^2\right] + \mathbb{E}_{\mathcal{D}}\left[(\bar{y}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x}))^2\right] + 2\mathbb{E}_{\mathcal{D},y|\boldsymbol{x}}\left[(y - \bar{y}(\boldsymbol{x}))(\bar{y}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x}))\right]
\end{aligned}$$



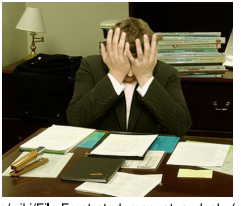
https://commons.wikimedia.org/wiki/File:Frustrated_man_at_a_desk_(cropped).jpg

The Bias-Variance decomposition

$$\mathbb{E}_{\mathcal{D},y|\boldsymbol{x}}\left[\left(y - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] = \mathbb{E}_{y|\boldsymbol{x}}\left[\left(y - \bar{y}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_{\mathcal{D}}\left[\left(\bar{y}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right]$$

$$\bar{f}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]$$

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\left(\bar{y}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\left(\bar{y}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x}) + \bar{f}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\left(\bar{y}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_{\mathcal{D}}\left[\left(\bar{f}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] + 2\mathbb{E}_{\mathcal{D}}\left[\left(\bar{y}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x})\right)\right]
\end{aligned}$$



https://commons.wikimedia.org/wiki/File:Frustrated_man_at_a_desk_(cropped).jpg

The Bias-Variance decomposition

$$\mathbb{E}_{\mathcal{D},y|\boldsymbol{x}}\left[\left(y - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] = \mathbb{E}_{y|\boldsymbol{x}}\left[\left(y - \bar{y}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_{\mathcal{D}}\left[\left(\bar{y}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right]$$

$$\bar{f}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]$$

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\left(\bar{y}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\left(\bar{y}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x}) + \bar{f}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\left(\bar{y}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_{\mathcal{D}}\left[\left(\bar{f}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] + 2\mathbb{E}_{\mathcal{D}}\left[\left(\bar{y}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x})\right)\right]
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D},y|\boldsymbol{x}}\left[\left(y - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] \\
&= \mathbb{E}_{y|\boldsymbol{x}}\left[\left(y - \bar{y}(\boldsymbol{x})\right)^2\right] + \left(\bar{y}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\right)^2 + \mathbb{E}_{\mathcal{D}}\left[\left(\bar{f}(\boldsymbol{x}) - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] \\
&= \mathrm{Var}_{y|\boldsymbol{x}}\left[y\right] + \left(\bar{y}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\right)^2 + \mathrm{Var}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]
\end{aligned}$$

The Bias-Variance decomposition

$$\mathbb{E}_{\mathcal{D}}\left[E^{\text{gen}}\right] = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D},y|\boldsymbol{x}}\left[\left(y - f_{\mathcal{D}}(\boldsymbol{x})\right)^2\right]\right]$$

$$\mathbb{E}_{\mathcal{D}}\left[E^{\text{gen}}\right] = \mathbb{E}_{\boldsymbol{x}}\left[\text{Var}_{y|\boldsymbol{x}}\left[y\right] + \left(\bar{y}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\right)^2 + \text{Var}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]\right]$$

https://commons.wikimedia.org/wiki/File:Frustrated_man_at_a_desk_(cropped).jpg

The Bias-Variance decomposition

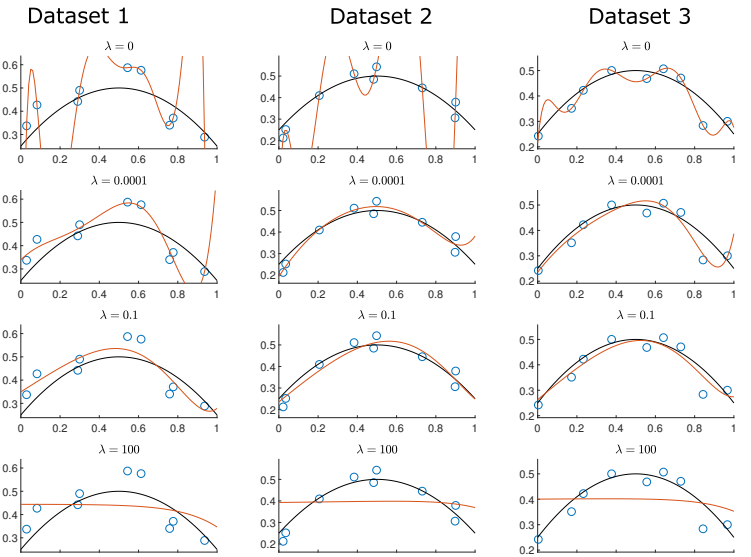
$$\mathbb{E}_{\mathcal{D}}\left[E^{\text{gen}}\right] = \mathbb{E}_{\boldsymbol{x}}\left[\text{Var}_{y|\boldsymbol{x}}\left[y\right] + \left(\bar{y}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\right)^2 + \text{Var}_{\mathcal{D}}\left[f_{\mathcal{D}}(\boldsymbol{x})\right]\right]$$

The first term does not depend at all upon our choice of model but simply represents the intrinsic difficulty of the problem. We cannot make this term any larger or smaller by selecting one model over another.

The second term is the **bias** term. It tells us how much the average values of models trained on different training datasets differ compared to the true mean of the data.

The third term is the **variance** term. It tells us how much the model wiggles when trained on different sets of training data. That is, when you train the models on N different (random) sets of training data and the models (the prediction curves) are nearly the same this term is small.

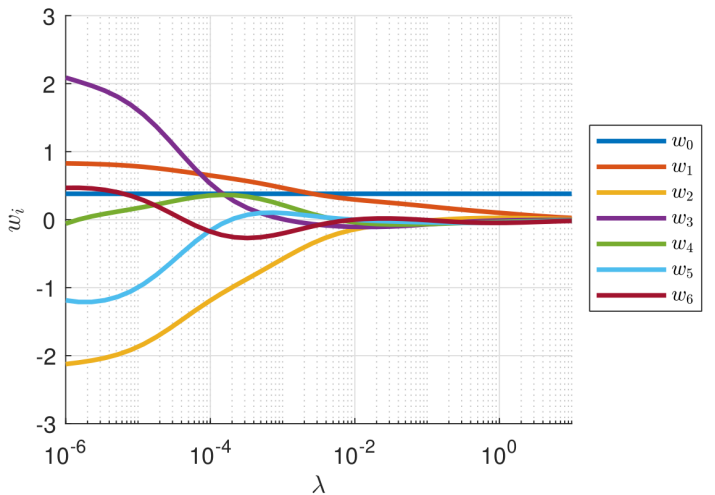
The bias variance decomposition



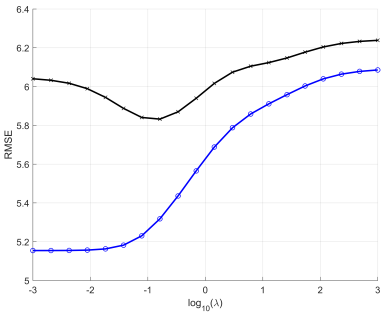
By regularization we can tradeoff bias and variance, in particular, we can hope to substantially reduce variance without introducing too much bias!

Parameters w^* as function of λ

$$E_\lambda(\boldsymbol{w}) = \sum_{i=1}^{N} (\hat{y}_i - w_0 - \hat{\boldsymbol{x}}_i{}^\top \boldsymbol{w})^2 + \lambda \|\boldsymbol{w}\|^2$$



Quiz 1, Bias-variance (Fall 2017)



Using 54 observations of a dataset about Basketball, we would like to predict the average points scored per game (y) based on the four features. For this purpose we consider regularized least squares regression which minimizes with respect to $\boldsymbol{w}$ the following cost function:

$$E(\boldsymbol{w}) = \sum_n (y_n - [1 \; x_{n1} \; x_{n2} \; x_{n3} \; x_{n4}]\boldsymbol{w})^2 + \lambda \boldsymbol{w}^\top \boldsymbol{w},$$

We consider 20 different values of λ and use leave-one-out cross-validation to estimate the performance (measured by mean-squared error) of each of these different values of λ and plot the result in the figure. For the value of $\lambda = 0.6952$ the following model is identified:

$$f(\boldsymbol{x}) = 2.76 - 0.37x_1 + 0.01x_2 + 7.67x_3 + 7.67x_4.$$

Which one of the following statements is correct?

A. In the figure the blue curve with circles corresponds to the training error whereas the black curve with crosses corresponds to the test error.

B. According to the model defined for $\lambda = 0.6952$ increasing a players height x_1 will increase his average points scored per game.

C. There is no optimal way of choosing λ since increasing λ reduces the variance but increases the bias.

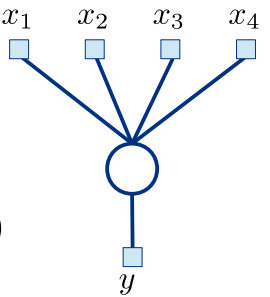
D. As we increase λ the 2-norm of the weight vector $\boldsymbol{w}$ will also increase.

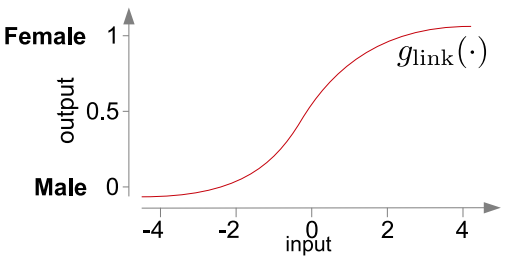
E. Don't know.

The correct answer is A: The blue curve monotonically increases with λ reflecting a worse fit to the training set as we increase λ using regularization we can reduce the variance by introducing bias and the black curve indicates that an optimal tradeoff at around $10^{-0.8}$ as reflected by the test error indicated in the black curve being minimal. As we increase λ we will penalize the weights according to the squared 2-norm more and more and thus the 2-norm will be reduced. Finally, according to the fitted model we observe that the coefficient in front of x_1 (Height) is negative thus indicating that an increase in height will reduce the models prediction of average points scored per game.

General linear model

- **Remember the generalized linear model?**
  - Data $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$
  - Model $f(\boldsymbol{x}) = g_{\text{link}}(\boldsymbol{x}^\top \boldsymbol{w})$
  - Cost function $d\big(y, f(\boldsymbol{x})\big)$
  - Parameters $\boldsymbol{w} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} d\big(y_n, f(\boldsymbol{x}_n)\big)$

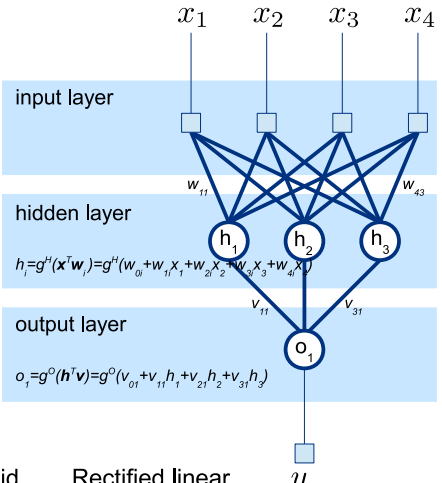




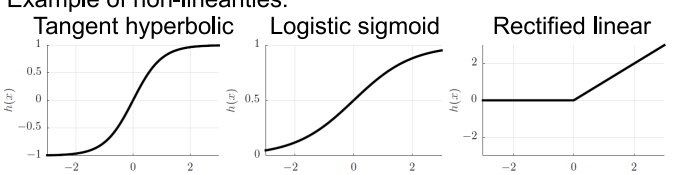
Artificial neural networks

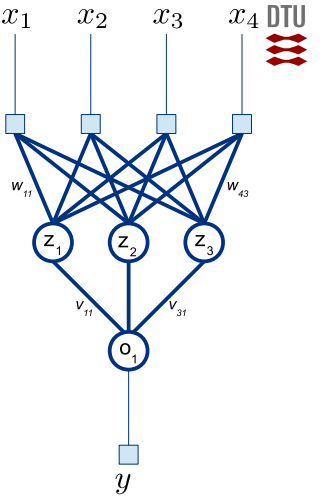
Feed forward network

- Each "neuron"
  - Computes a non-linear function of the sum of its inputs
  - Is just like a generalized linear model
  - Has its own set of parameters
- Modeling choices
  - Cost function
  - Non-linearities
  - Number of neurons and hidden layers
  - Selection of inputs
- Parameter estimation using numerical optimization methods
- Very flexible model: Can easily overfit



Example of non-linearities:





Data: $\{\boldsymbol{x}_i, y_i\}$

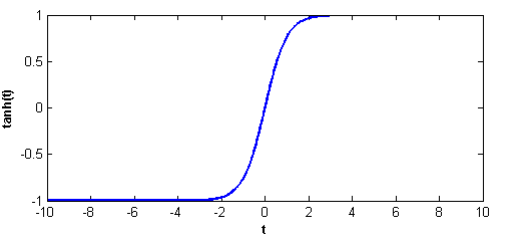
Model: $f(\boldsymbol{x}) = h^{(2)}\left(v_{10} + \sum_{j=1}^{H} v_{1j} h^{(1)}\left(\tilde{\boldsymbol{x}}^\top \boldsymbol{w}_j\right)\right)$

Distance: $d(y, f(\boldsymbol{x}))$

Cost: $E = \sum_{i=1}^{N} d(y_i, f(\boldsymbol{x}_i))$

Common choices

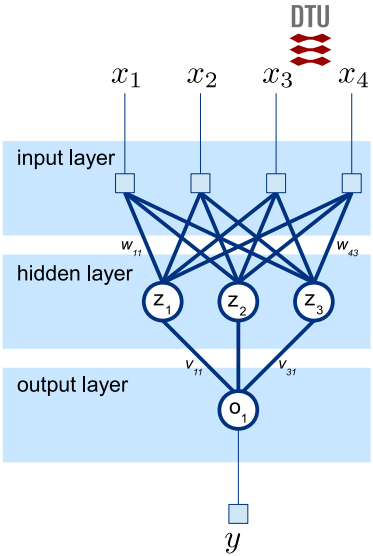
$$h^{(1)}(x) = \tanh(x)$$
$$h^{(2)}(x) = x$$
$$d(y, f(\boldsymbol{x})) = (y - f(\boldsymbol{x}))^2$$



Neurons and layers

Recall:

$$f(\boldsymbol{x}) = h^{(2)}\left(v_{10} + \sum_{j=1}^{H} v_{1j} h^{(1)}\left(\tilde{\boldsymbol{x}}^\top \boldsymbol{w}_j\right)\right)$$



- Let $z_j^{(1)}$ be output of j'th hidden unit

$$z_j^{(1)} = h^{(1)}\left(\boldsymbol{w}_j^{(1)\top} \tilde{x}\right)$$

Abbreviated $\boldsymbol{z}^{(1)} = h^{(1)}\left(\boldsymbol{W}^{(1)}\tilde{x}\right)$

- Output

$$f(\boldsymbol{x}) = h^{(2)}\left(v_{10} + \sum_{j=1}^{H} v_{1j} z_j^{(1)}\right) = h^{(2)}\left(\boldsymbol{W}^{(2)}\tilde{\boldsymbol{z}}^{(1)}\right)$$

We consider each $z_j^{(1)}$ a neuron and $\boldsymbol{z}^{(1)}$ a (hidden) layer

Quiz 2, Artificial Neural Network (Fall 2017)

We will consider an artificial neural network (ANN) trained to predict the average score of a player (i.e., y). The ANN is based on the model:

$$f(\boldsymbol{x}, \boldsymbol{w}) = w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} h^{(1)}([1\ \boldsymbol{x}]\boldsymbol{w}_j^{(1)}).$$

where $h^{(1)}(x) = \max(x, 0)$ is the rectified linear function used as activation function in the hidden layer (i.e., positive values are returned and negative values are set to zero). We will consider an ANN with two hidden units in the hidden layer defined by:

$$\boldsymbol{w}_1^{(1)} = \begin{bmatrix} 21.78 \\ -1.65 \\ 0 \\ -13.26 \\ -8.46 \end{bmatrix}, \boldsymbol{w}_2^{(1)} = \begin{bmatrix} -9.60 \\ -0.44 \\ 0.01 \\ 14.54 \\ 9.50 \end{bmatrix},$$

and $w_0^{(2)} = 2.84$, $w_1^{(2)} = 3.25$, and $w_2^{(2)} = 3.46$.
What is the predicted average score of a basketball player with observation vector $\boldsymbol{x}^* = [6.8\ 225\ 0.44\ 0.68]$?

A. 1.00

B. 3.74

C. 8.21

D. 11.54

E. Don't know.

The output is given by:

$$
\begin{aligned}
f(\boldsymbol{x}, \boldsymbol{w}) &= 2.84 \\
&\quad + 3.25 \cdot \max([1 \ 6.8 \ 225 \ 0.44 \ 0.68] \cdot \begin{bmatrix} 21.78 \\ -1.65 \\ 0 \\ -13.26 \\ -8.46 \end{bmatrix}, 0) \\
&\quad + 3.46 \max([1 \ 6.8 \ 225 \ 0.44 \ 0.68] \cdot \begin{bmatrix} -9.60 \\ -0.44 \\ 0.01 \\ 14.54 \\ 9.50 \end{bmatrix}, 0) \\
&= 2.84 + 3.25 \cdot \max(-1.027, 0) + 3.46 \max(2.516, 0) \\
&= 11.54
\end{aligned}
$$

Generalization 1: Multiple outputs

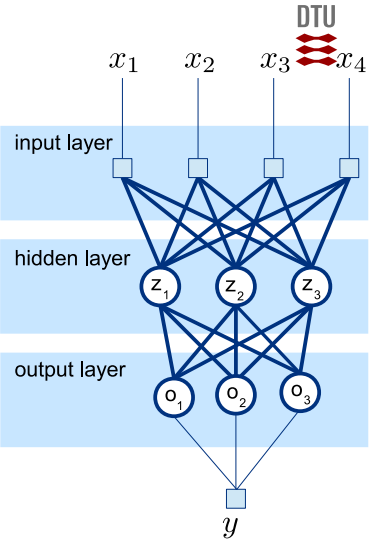
- As before define: $\boldsymbol{z}^{(1)} = h^{(1)}\left(\boldsymbol{W}^{(1)}\tilde{x}\right)$
- Now let $\boldsymbol{W}^{(2)}$ be a $C \times H$ matrix then:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) = h^{(2)}\left(\boldsymbol{W}^{(2)}\tilde{\boldsymbol{z}}^{(1)}\right)$$

  will be C-dimensional

- Re-define error function

$$E = \sum_{i=1}^{N} \|\boldsymbol{y}_i - \boldsymbol{f}(\boldsymbol{x}_i)\|_2^2$$



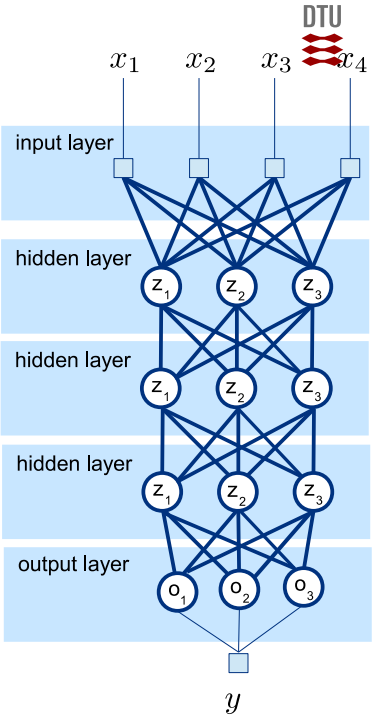
Generalization 2: Multiple layers

- Define $\boldsymbol{z}^{(0)} = \boldsymbol{x}$
- For each layer $l = 1, \ldots, L$ compute

$$z_j^{(l)} = h^{(l)}\left(\boldsymbol{W}^{(l)}\tilde{\boldsymbol{z}}^{(l-1)}\right)$$

- Output is simply

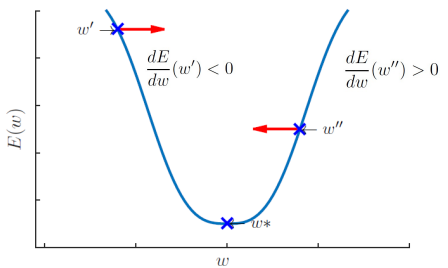
$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{z}^{(L)}$$

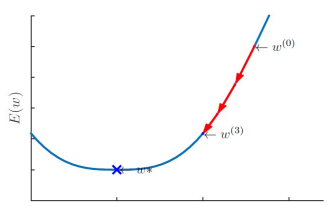


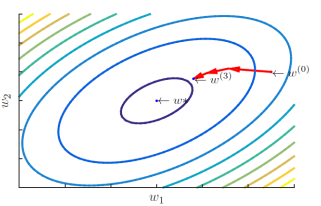
Gradient descent

- Start from an initial guess at $\boldsymbol{w}^*$, $\boldsymbol{w}^{(0)}$
- At step t of the algorithm, modify $\boldsymbol{w}^{(t-1)}$ to produce a better guess $\boldsymbol{w}^{(t)}$:

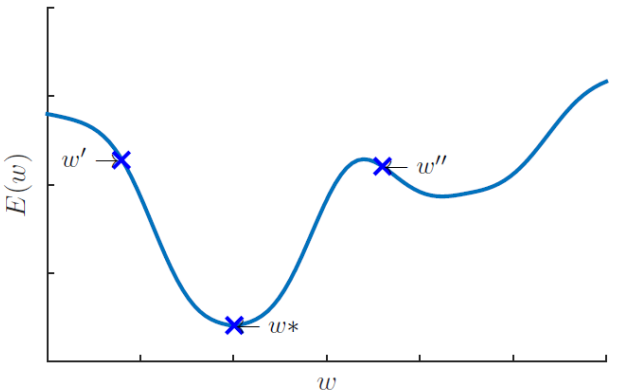
$$\boldsymbol{w}^{(t)} = \boldsymbol{w}^{(t-1)} - \epsilon \frac{dE}{d\boldsymbol{w}}(\boldsymbol{w}^{(t-1)})$$



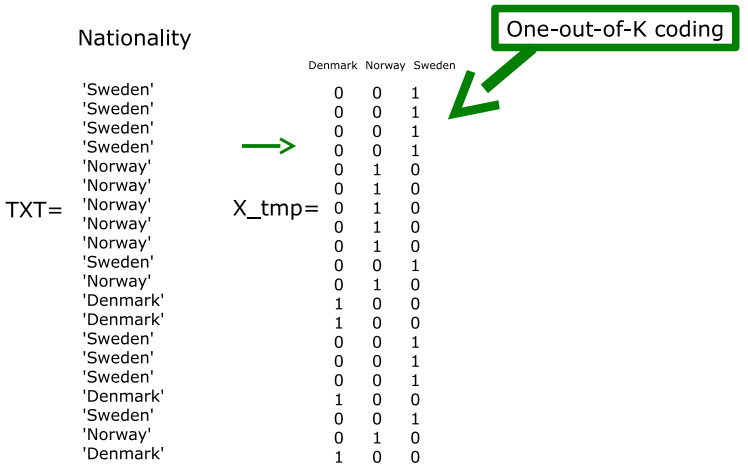




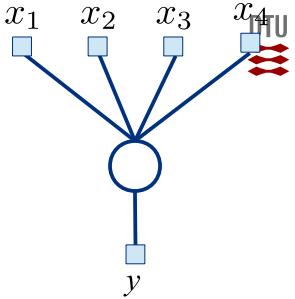
Contrary to least-squares linear regression and logistic regression ANNs have issues of local minima



Single and multi-class: One out of K coding



Multi-class classification



- Logistic regression, $y = 0, 1$:

$$p(y|\theta) = \theta^y (1-\theta)^{1-y}$$
$$\theta = \sigma(\boldsymbol{x}^\top \boldsymbol{w})$$

- Multinomial regression, $y = 1, 2, \ldots, K$

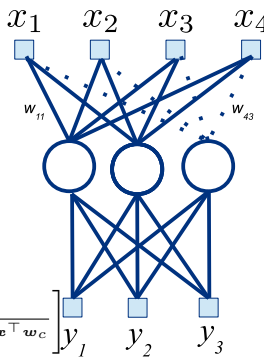
z_k : one-of-K encoding of y,

$$p(y|\boldsymbol{\theta}) = \prod_{i=1}^{K} \theta_k^{z_k}$$

$$\boldsymbol{\theta} = \text{softmax}\left(\begin{bmatrix}\boldsymbol{x}^\top \boldsymbol{w}_1 & \cdots & \boldsymbol{x}^\top \boldsymbol{w}_K\end{bmatrix}\right)$$

$$= \begin{bmatrix} \frac{e^{\boldsymbol{x}^\top \boldsymbol{w}_1}}{\sum_{c=1}^{K} e^{\boldsymbol{x}^\top \boldsymbol{w}_c}} & \cdots & \frac{e^{\boldsymbol{x}^\top \boldsymbol{w}_{K-1}}}{\sum_{c=1}^{K} e^{\boldsymbol{x}^\top \boldsymbol{w}_c}} & \frac{e^{\boldsymbol{x}^\top \boldsymbol{w}_K}}{\sum_{c=1}^{K} e^{\boldsymbol{x}^\top \boldsymbol{w}_c}} \end{bmatrix}$$

or: $\boldsymbol{\theta} = \begin{bmatrix} \frac{e^{\boldsymbol{x}^\top \boldsymbol{w}_1}}{1+\sum_{c=1}^{K-1} e^{\boldsymbol{x}^\top \boldsymbol{w}_c}} & \cdots & \frac{e^{\boldsymbol{x}^\top \boldsymbol{w}_{K-1}}}{1+\sum_{c=1}^{K-1} e^{\boldsymbol{x}^\top \boldsymbol{w}_c}} & \frac{1}{1+\sum_{c=1}^{K-1} e^{\boldsymbol{x}^\top \boldsymbol{w}_c}} \end{bmatrix}$



Connection to neural networks

Multinomial regression:

- Define:

$$\boldsymbol{\theta} = \left[\frac{e^{\boldsymbol{x}^\top \boldsymbol{w}_1}}{1+\sum_{c=1}^{K-1} e^{\boldsymbol{x}^\top \boldsymbol{w}_c}} \quad \cdots \quad \frac{1}{1+\sum_{c=1}^{K-1} e^{\boldsymbol{x}^\top \boldsymbol{w}_c}} \right]$$

- Cost function is ($z_{i\cdot}$ is one-of-K encoding of y_i)

$$E = -\sum_{i=1}^{N} \log p(y_i|\boldsymbol{x}_i) = -\sum_{i=1}^{N}\sum_{c=1}^{K} z_{ic} \log \theta_{ic}$$

Multi-class neural network:

- Suppose $\tilde{y}_1, \ldots, \tilde{y}_K$ are outputs of a neural network
- Define

$$\boldsymbol{\theta} = \left[\frac{e^{\tilde{\boldsymbol{y}}^\top \boldsymbol{w}_1}}{\sum_{c=1}^{K} e^{\tilde{\boldsymbol{y}}^\top \boldsymbol{w}_c}} \quad \cdots \quad \frac{e^{\tilde{\boldsymbol{y}}^\top \boldsymbol{w}_K}}{\sum_{c=1}^{K} e^{\tilde{\boldsymbol{y}}^\top \boldsymbol{w}_c}} \right]$$

- Cost function is:

$$E = -\sum_{i=1}^{N} \log p(y_i|\tilde{\boldsymbol{y}}_i) = -\sum_{i=1}^{N}\sum_{c=1}^{K} z_{ic} \log \theta_{ic}$$

Quiz 3, Multinomial Regression (Spring 2016)

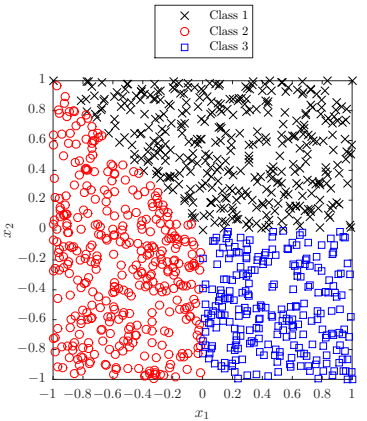


Figure 1: Observations labelled with the most probable class

Consider a multinomial regression classifier for a three-class problem where for each point $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top$ we compute the class-probability using the softmax function

$$P(\hat{y} = k) = \frac{e^{\boldsymbol{w}_k^\top \boldsymbol{x}}}{e^{\boldsymbol{w}_1^\top \boldsymbol{x}} + e^{\boldsymbol{w}_2^\top \boldsymbol{x}} + e^{\boldsymbol{w}_3^\top \boldsymbol{x}}}.$$

A dataset of $N = 1000$ points where each point is labeled according to the maximum class-probability is shown in Figure 1. Which setting of the weights was used?

A. $w_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $w_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

B. $w_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $w_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

C. $w_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $w_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

D. $w_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $w_3 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

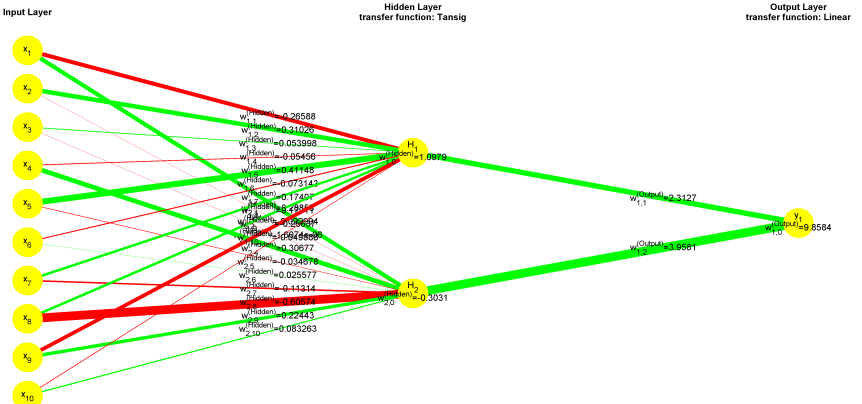
E. Don't know.

Consider for instance the point $\boldsymbol{x}$ where $x_1 = 0$ and $x_2 = 1$. Then, letting $y_k = \boldsymbol{w}_k^T \boldsymbol{x}$, we obtain:

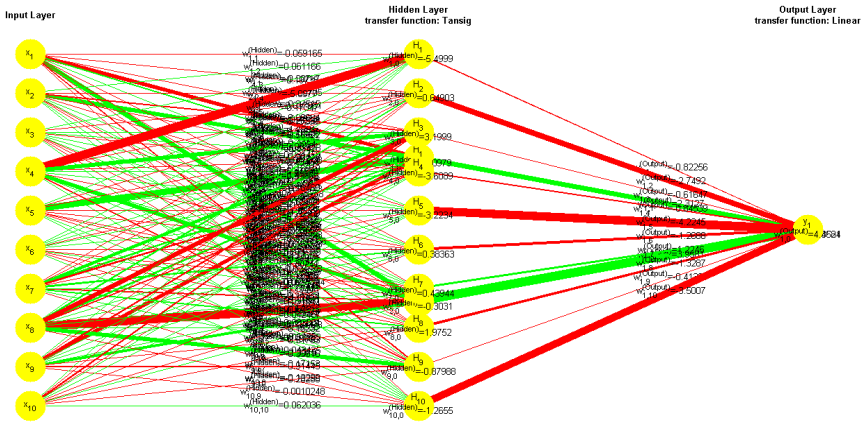
$$
\begin{aligned}
A &: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & -1 \end{bmatrix} \\
B &: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} -1 & -1 & 1 \end{bmatrix} \\
C &: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 \end{bmatrix} \\
D &: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 1 \end{bmatrix}
\end{aligned}
$$

Next, since the multinomial regression function preserves order we need only consider the maximal value. Accordingly the point $\boldsymbol{x}$ is only classified to the correct class 1 for option C.

Interpreting neural networks can be difficult



Interpreting neural networks can be difficult



Resources

https://www.youtube.com Exellent video resource explaining the concepts behind neural networks (https://www.youtube.com/watch?v=aircAruvnKk&list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi)

http://playground.tensorflow.org Sleek interactive neural network example where you can examine the effect of different number of hidden neurons, activation functions, and many other things on training (http://playground.tensorflow.org/)

https://www.tensorflow.org Most popular and well-documented deep learning framework. While well documented, notice it requires some python knowledge (https://www.tensorflow.org/)

https://pytorch.org Upcoming (and in some ways slightly simpler) framework for deep learning; alternative to tensorflow (https://pytorch.org/)

Mid-term quiz 1

In the analysis of house prices the following attributes were collected for a house: The year the house was built (denoted YEAR), the size of the house given in square meters (denoted SIZE) the county in which the house is located (denoted LOCATION). Which statement about the three attributes is correct?

A. YEAR is ratio, SIZE is interval and LOCATION is nominal

B. YEAR is interval, SIZE is ratio and LOCATION is nominal

C. YEAR is interval, SIZE is ratio and LOCATION is ordinal

D. YEAR is interval, SIZE is ratio and LOCATION is interval

E. Don't know.

Year data types do not have a zero with a physical meaning and are therefore interval. Size has a physically relevant zero and is therefore ratio. Meanwhile, location is just an identifier which only support similarity-comparison and is therefore nominal.

Mid-term quiz 2

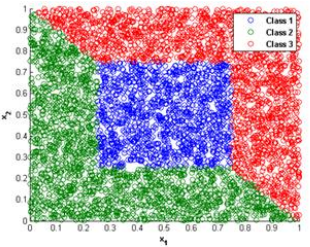
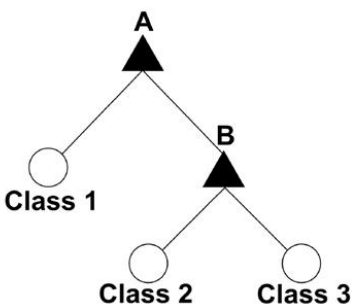


Figure 1



Consider the classification problem given in figure 1 and the Decision Tree shown below it with two decision nodes denoted A and B. We will let $\boldsymbol{x}_n = (x, y)$ denote a 2-dimensional observation such that $\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}$ denotes the subtraction of 0.5 from each of the two coordinates of $\boldsymbol{x}_n$. Which one of the following classification rules would lead to a correct classification of the data?

A. A: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_1 \leq 0.25$, B : $\|\boldsymbol{x}_n\|_\infty \leq 1$

B. A: $\|\boldsymbol{x}_n\|_1 \leq 1$, B: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq \infty$

C. A: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq 0.25$, B : $\|\boldsymbol{x}_n\|_\infty \leq 1$

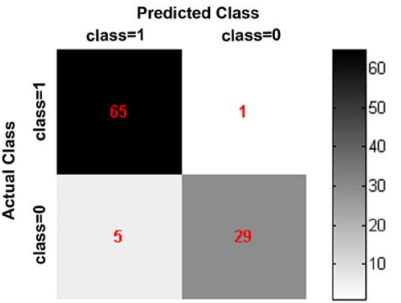
D. A: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_\infty \leq 0.25$, B : $\|\boldsymbol{x}_n\|_1 \leq 1$

E. Don't know.

The right answer is D. Recall the shape associated with the different L_p-norms: $p = 2$ is a circle (Euclidean distance), $p = \infty$ a square, and $p = 1$ a square rotated 45 degrees. If we therefore first as at A if the observation is within the square (if yes, classify as the blue class, otherwise go to next split) and then at the next split ask if it within the L_1-norm of origo (distance 1), we get the diagonal decision boundary. This can be implemented by option D.

Mid-term quiz 3

A classifier has the confusion matrix given in the figure below. Which statement about the classifier is correct?



A. The Accuracy is 94% and the Error rate is 6%

B. The Accuracy is 6% and the Error rate is 94%

C. The Accuracy is 65% and the Error rate is 35%

D. There is insufficient information in the confusion matrix to determine the Accuracy and Error rate.

E. Don't know.

Accuracy is total number of correct choices divided by total number of observations. Therefore, the accuracy is $\frac{65+29}{6+65+29} = \frac{94}{100}$ or 95%. The right answer is therefore A.

Mid-term quiz 4

Which statement about crossvalidation is wrong?

A. Cross-validation can be used to estimate the generalization error.

B. Leave one out cross-validation is more computationally expensive than 10 fold crossvalidation.

C. Holding out one third of the data for validation is faster but less accurate than performing 10 fold cross-validation.

D. The same test set can be used for model selection as well as evaluation of the generalization performance of the model.

E. Don't know.

The last option D is wrong because if we both select a model on a test set and then later use it for estimating the generalization error we will not obtain an unbiased estimate of the generalization error since we have already tuned the model on the test set. For this task, one should use two-layer CV.

Mid-term quiz 5

Consider a data set of four features: A, B, C, and D that are applied in a classification algorithm. The table below shows the cross-validated Error rate when using different combinations of the features.

Feature(s)	Error rate
A	0.40
B	0.45
C	0.33
D	0.42
A and B	0.20
A and C	0.25
A and D	0.34
B and C	0.29
B and D	0.42
C and D	0.40
A and B and C	0.13
A and B and D	0.17
B and C and D	0.10
A and C and D	0.15
A and B and C and D	0.28

We will apply a forward feature selection algorithm. Which feature set will the selection algorithm choose?

A. C

B. B and C and D

C. A and B

D. A and B and C

E. Don't know.

Forward selection will attempt to minimize the error rate. It will first select C, then select lowest of the next options containing C, i.e. A, C, and then A, B, C. Therefore, option D is correct.

Mid-term quiz 6

When training a decision tree we will use the classification error as impurity measure $I(t)$ given by $I(t) = 1 - \max_i[p(i|t)]$ where $p(i|t)$ denotes the fraction of data objects belonging to class i at a given node t. We will use Huntâ€™s algorithm to grow the tree and recall that the purity gain is given by:

$$\Delta = I(\text{ Parent }) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

where N is the total number of data objects at the parent node, k is the number of child nodes and $N(v_j)$ is the number of data objects associated with the child node, v_j. We will consider classification of Iris flowers into Iris-Setosa, Iris-Virginica and Iris-Versicolor. At a potential split we have:

- Before the split: 5 Iris-Setosa, 10 Iris-Virginica and 10 Iris Versicolor.

After the split

- 0 Iris-Setosa, 8 Iris-Virginica and 2 Iris-Versicolor in the left node.
- 5 Iris-Setosa, 2 Iris-Viriginica and 8 Iris-Versicolor in the right node.

Which statement is correct?

A. The purity gain is $\Delta = \frac{3}{5}$

B. The purity gain is $\Delta = \frac{3}{15}$

C. The purity gain is $\Delta = \frac{6}{25}$

D. The purity gain is $\Delta = \frac{7}{15}$

E. Don't know.

There are a total of $N = 25$ observations and the number in the two branches are $N_1 = 10$ and $N_2 = 15$. In the base branch, the maximum class-probability is $\frac{10}{25}$ and so $I_0 = 1 - \frac{10}{25} = \frac{15}{25} = \frac{3}{5}$. Similarly, we compute $I_1 = \frac{1}{5}$ and $I_2 = 1 - \frac{8}{15} = \frac{7}{15}$. We now have

$$\Delta = I_0 - \frac{N_1}{N} I_1 - \frac{N_2}{N} I_2 = \frac{3}{5} - \frac{10}{25}\frac{1}{5} - \frac{15}{25}\frac{7}{15} \quad (1)$$

$$= \frac{3}{5} - 225 - \frac{7}{25} = \frac{15 - 2 - 7}{25} = \frac{6}{25} \quad (2)$$

or C.

Mid-term quiz 7

When people are well rested and take an exam their chance of passing the exam is 90%, however, when people are not well rested there chance of passing the exam is only 40%. On any given day 80% of people are well-rested. What is the chance that a person passing the test is well rested?

A. $\frac{4}{10}$

B. $\frac{8}{10}$

C. $\frac{9}{10}$

D. $\frac{10}{11}$

E. Don't know.

Let R be rested and P be passing. Then the answer is

$$P(R|P) = \frac{P(P|R)P(R)}{P(P|\overline{R})P(\overline{R}) + P(P|R)P(R)} \quad (1)$$

$$= \frac{0.9 \times 0.8}{0.4 \times 0.2 + 0.9 \times 0.8} \quad (2)$$

$$= \frac{0.9}{0.1 + 0.9} = \frac{9}{1 + 9} = \frac{9}{10} \quad (3)$$

and so C is correct.

Mid-term quiz 8

When carrying out a principal component analysis of a dataset with four attributes we obtain the following singular values $\sigma_1 = 4$, $\sigma_2 = 2$, $\sigma_3 = 1$, and $\sigma_4 = 0$.

Which one of the following statements is wrong?

A. The first principal component accounts for more than 60% of the variation in the data.

B. The third principal component accounts for less than 5% of the variation in the data.

C. The second principal component accounts for more than 20% of the variation in the data.

D. The data can be perfectly represented in a three dimensional sub-space.

E. Don't know.

The variance explained of a given coordinate is $\frac{\sigma_i^2}{\sum_{i=1}^4 \sigma_i^2}$. Therefore, the variance explained by the second coordinate is $\frac{4}{21} < \frac{1}{5}$ and so C is the right answer.

Mid-term quiz 9

Consider the following sequence of numbers

$$x = \begin{bmatrix} 0 & 1 & 1 & 1 & 2 & 3 & 4 & 4 & 5 & 14 \end{bmatrix}.$$

What is the sum of the mean, the median and the mode of these numbers, i.e. what is the value: $y =$ $\text{mean}(x) + \text{median}(x) + \text{mode}(x)$?

A. $y = 1$

B. $y = 6$

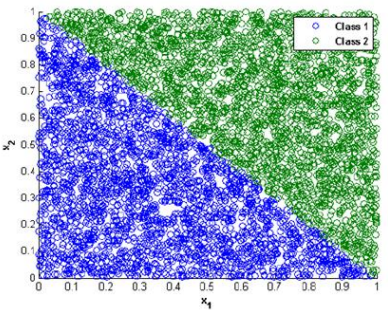
C. $y = 7$

D. $y = 11$

E. Don't know.

The mode is 1 (most common number). The median is 2.5 (since the list is ordered and contains an even number of elements it is the average of 2 and 3) and the mean is sum divided by 10 or 3.5. Therefore, the answer is 7.

Mid-term quiz 10



Consider the classification problem given in the figure below where x_1 and x_2 are used as features for a logistic regression classifier and a decision tree. The considered logistic regression models all include the constant term w0. Which one of the following statements is **wrong**?

A. The two classes can be perfectly separated by a logistic regression model using x_1 and x_2 as features.

B. A decision tree with less than five nodes, all of the usual axis-aligned form $x_1 > a$ or $x_2 > b$ for different values of a, b, can perfectly separate the classes using only x_1 and x_2 as features.

C. A logistic regression model can perfectly separate the two classes using only the feature z given by $z = x_1 + x_2$.

D. In logistic regression the probability that each observation belong to the two classes can be derived from the logistic function.

E. Don't know.

B: To see why B is wrong, note the decision boundary will of such a tree will consist of rectangles with axis-oriented sides. The other options are easily seen to be correct and for C, note that the boundary shown in the plot corresopnsd to $z > 1$.