02450 Introduction to Machine Learning and Data Mining

Week 8: Artificial Neural Networks and Bias/Variance

Georgios Arvanitidis

25 March 2025

DTU Compute, Technical University of Denmark

Today

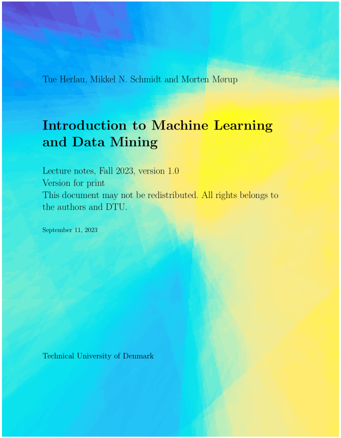
Feedback Groups of the day:

Benjamin Noah Lumbye, Asghar Ali Khan Menashi, Ditte Marie Nordlund Boye, Zhenlin Xie, Maiken Rosa Ellested, Lukas Tej Marner, Nil Mataró Llobet, Dario Cannistra, Zeeshawn Hasnain, Eleni-Sofia Tseperi, Louis Adrian Teufel, Martin Wurlitzer Romme, Pablo Bee Olmedo, Bence Göblyös, Alona Konstantinova, Jonatan Muxoll Larsen, Freja Egelund Grønnemose, Nanny Henriksen, Dina Bech Rindorf, Moira Josefine Losch, Phi Thanh Vo, Jakob Schwanenflügel, Oliver Colmer, Sophia Helena Andersen, Rehab Salaheddin Abu Rashed, Mads Tornby Christoffersen, Adam Ajane, Olivia Linh Merrild Knudsen, Dominik Brymora, Lisa Liv Braunstein Jonsson, Clara Marie Zacho Hansen, Elias Rajabi, Aurélien Paul Armand Cresp, Kim Viet Tran, Ingrid Helene Zimmermann Petersen, Andrea Albasini, Pauline Schielke, Mikkel Vinther Fritzel, Marina Gallego Jene, Elias Storm Vedel Jørgensen, Lucas Lydik Bessing, Nicholas Lentfer Tachibana Kristiansen

Reading/homework material:

Chapter 14, 15

P15.1, P15.2, P15.3



Tue Herlau, Mikkel N. Schmidt and Morten Mørup

Introduction to Machine Learning and Data Mining

Lecture notes, Fall 2023, version 1.0
Version for print
This document may not be redistributed. All rights belongs to the authors and DTU.

September 11, 2023

Technical University of Denmark

Lecture Schedule

1. Introduction
   4 February: C1,C2

Data: Feature extraction, and visualization

2. Summary statistics, similarity and visualization
   11 February: C4,C7
3. Computational linear algebra and PCA
   18 February: C3
4. Probability and probability densities
   25 February: C5, C6

Supervised learning: Classification and regression

5. Decision trees and linear regression
   4 March: C8, C9 **(Project 1 due 6 March at 17:00)**
6. Overfitting, cross-validation and Nearest Neighbor
   11 March: C10, C12
7. Performance evaluation, Bayes, and Naive Bayes
   18 March: C11, C13
8. **Artificial Neural Networks and Bias/Variance**
   25 March: C14, C15
9. AUC and ensemble methods
   1 April: C16, C17

Unsupervised learning: Clustering and density estimation

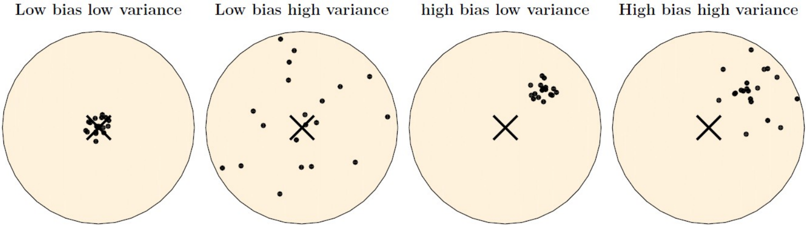
10. K-means and hierarchical clustering
    8 April: C18 **(Project 2 due 10 April at 17:00)**
11. Mixture models and density estimation
    22 April: C19, C20
12. Association mining
    29 April: C21

Recap

13. Recap and discussion of the exam
    6 May: C1-C21

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

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:

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

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

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

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

- 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

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

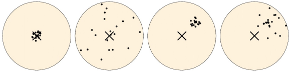
Selecting λ

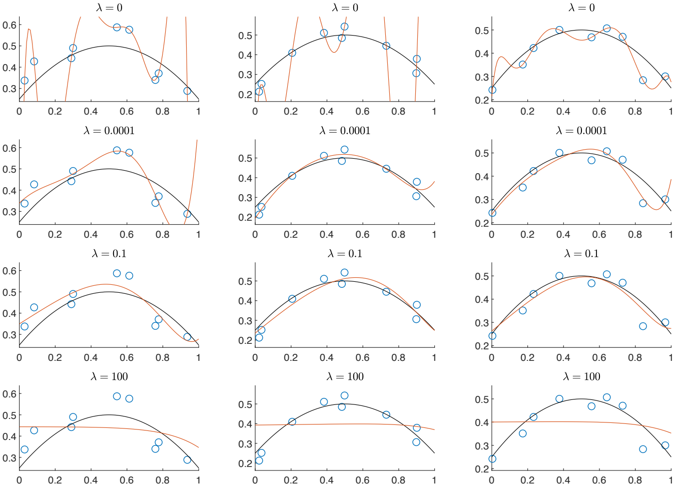
- Suppose

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

- 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)





Bias-variance decomposition

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

We first consider $\boldsymbol{x}$ fixed

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

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

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

Bias-variance decomposition

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

We first consider $\boldsymbol{x}$ fixed

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

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

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

The Bias-Variance decomposition

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

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

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

The Bias-Variance decomposition

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

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

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

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

The Bias-Variance decomposition

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

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

The Bias-Variance decomposition

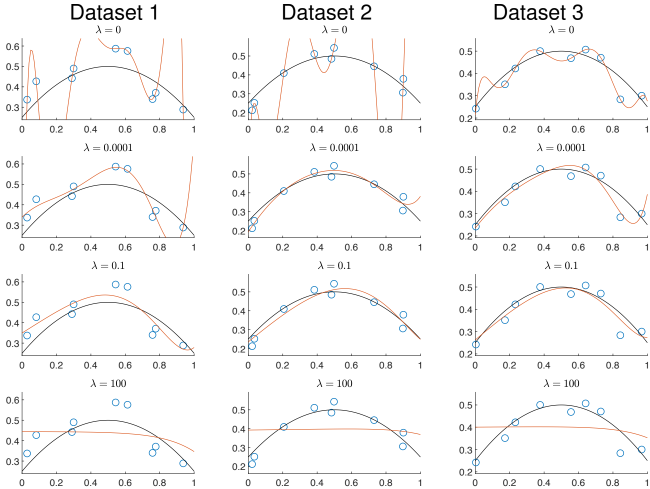
$$\mathbb{E}_{\mathcal{D}}[E^{gen}] = \mathbb{E}_{\boldsymbol{x}}\left[\underline{\mathsf{Var}_{y|\boldsymbol{x}}[y]} + \underline{(\bar{y}(\boldsymbol{x}) - \bar{f}(\boldsymbol{x}))^2} + \underline{\mathsf{Var}_{\mathcal{D}}[f_{\mathcal{D}}(\boldsymbol{x})]}\right]$$

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.

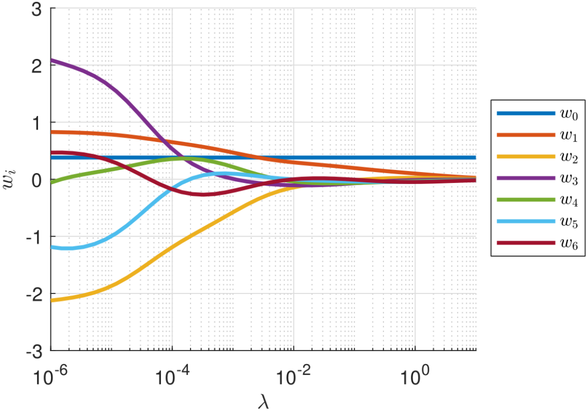
The bias variance decomposition



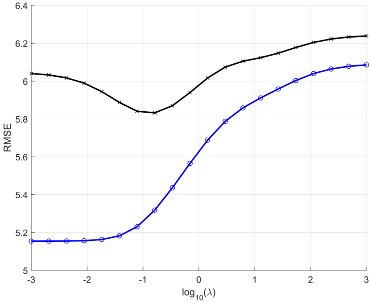
By regularization we can trade-off bias and variance, in particular, we can hope to substantially reduce variance without introducing too much bias!

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.

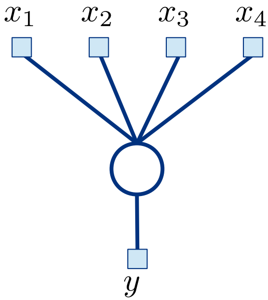
D. As we increase λ the 2-norm of the weight vector $\boldsymbol{w}$ will also increase.

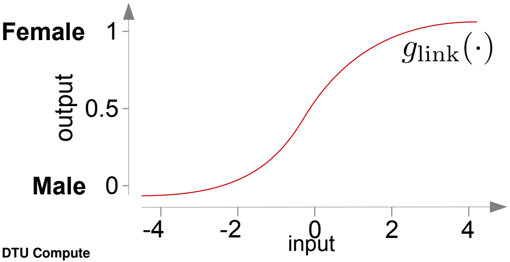
E. Don't know.

Generalized 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 $d(y, f(\boldsymbol{x}))$
- Parameters $\boldsymbol{w} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \sum_{n=1}^{N} d(y_n, f(\boldsymbol{x}_n))$

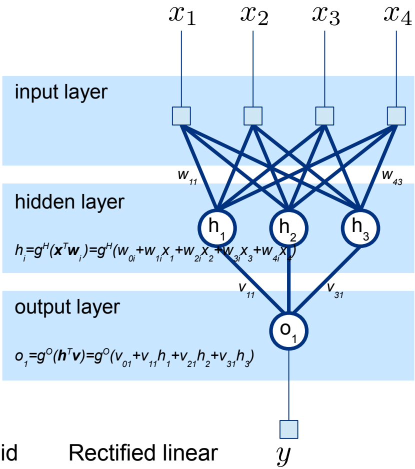




Artificial neural networks

Feed forward neural network

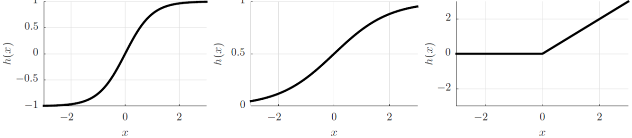
- Each “neuron”
  - Computes a nonlinear function of the sum of its inputs
  - Is just like a generalized linear model
  - Has its own set of parameters
- Modeling choices
  - Cost function
  - Non-linearities
  - Number of neurons and hidden layers
  - Selection of inputs
- Parameter estimation using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



$h_i=g^H(\mathbf{x}^T\mathbf{w}_i)=g^H(w_{0i}+w_{1i}x_1+w_{2i}x_2+w_{3i}x_3+w_{4i}x_4)$

$o_1=g^O(\mathbf{h}^T\mathbf{v})=g^O(v_{01}+v_{11}h_1+v_{21}h_2+v_{31}h_3)$

Example of non-linearities:


Tangent hyperbolic

Logistic sigmoid

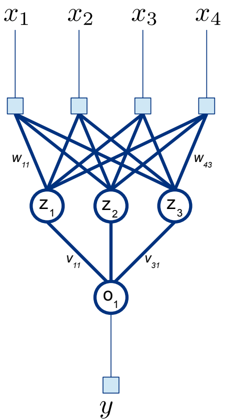
Rectified linear

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

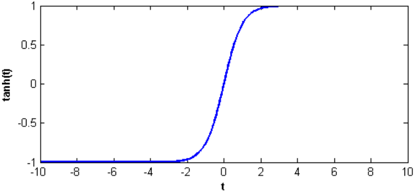


Common choices

$$h^{(1)}(x) = \tanh(x)$$

$$h^{(2)}(x) = x$$

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



Neurons and layers

Recall:

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

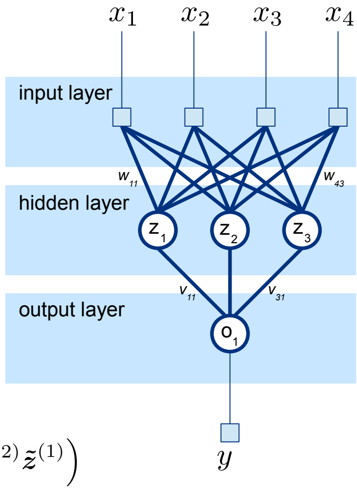
- Let $z_j^{(1)}$ be output of j'th hidden unit

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

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

- Output

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



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.

Generalization 1: Multiple outputs

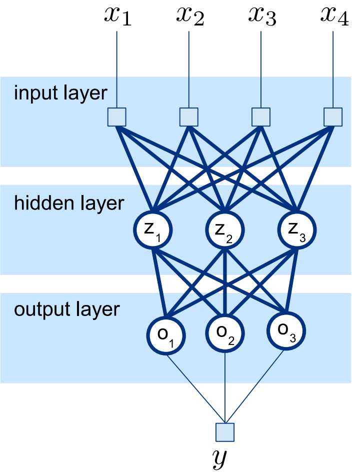
- As before define: $\boldsymbol{z}^{(1)} = h^{(1)}\left(\boldsymbol{W}^{(1)}\tilde{x}\right)$
- Now let $\boldsymbol{W}^{(2)}$ be a $C \times H$ matrix then:

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

  will be C-dimensional

- Re-define error function

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



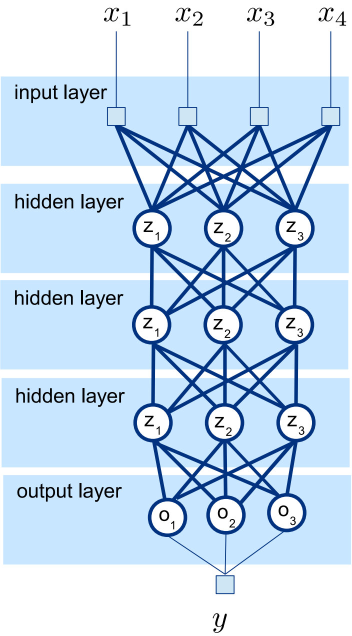
Generalization 2: Multiple layers

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

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

- Output is simply

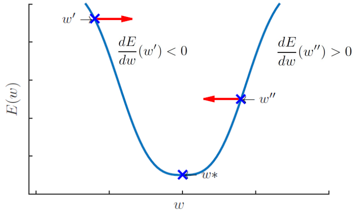
$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{z}^{(L)}$$

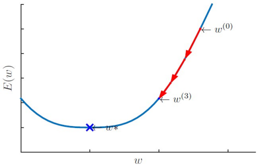


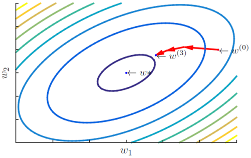
Gradient descent

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

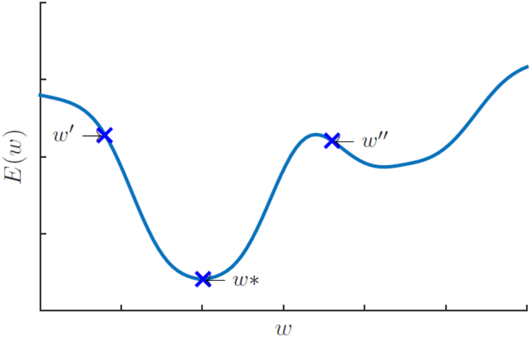
$$\boldsymbol{w}^{(t)} = \boldsymbol{w}^{(t-1)} - \varepsilon \frac{dE}{d\boldsymbol{w}}(\boldsymbol{w}^{(t-1)})$$







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



Single and multi-class: One out of K coding

Nationality

```
TXT=  'Sweden'
      'Sweden'
      'Sweden'
      'Sweden'
      'Norway'
      'Norway'
      'Norway'
      'Norway'
      'Norway'
      'Sweden'
      'Norway'
      'Denmark'
      'Denmark'
      'Sweden'
      'Sweden'
      'Sweden'
      'Denmark'
      'Sweden'
      'Norway'
      'Denmark'
```

$\longrightarrow$

$$\text{X\_tmp}= \begin{array}{ccc} \text{Denmark} & \text{Norway} & \text{Sweden} \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{array}$$

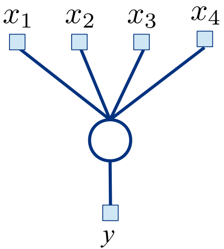
One-out-of-K coding

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, $y = 1, 2, \ldots, K$

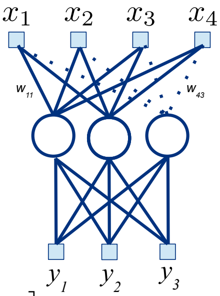
z_k : one-of-K encoding of y,

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

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

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

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)

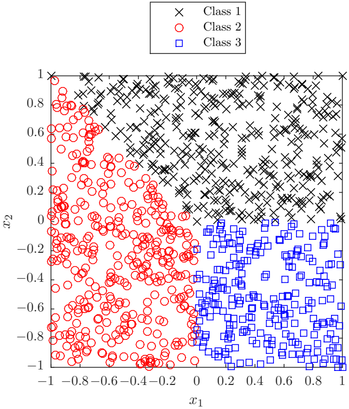


Figure 1: Observations labelled with the most probable class

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

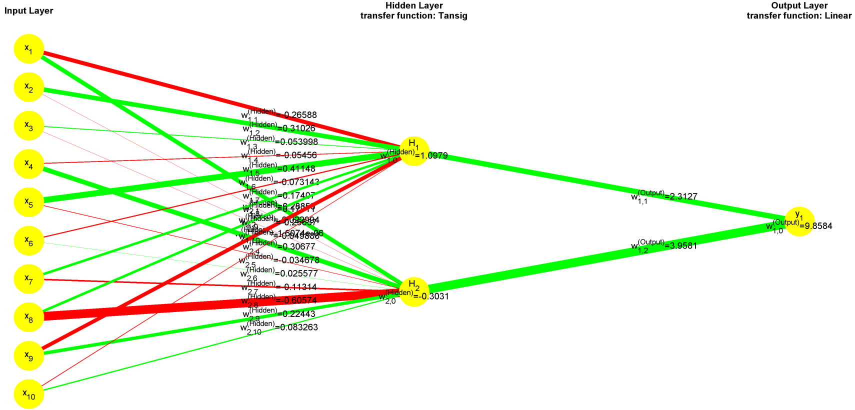
B. $w_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, w_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, w_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

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

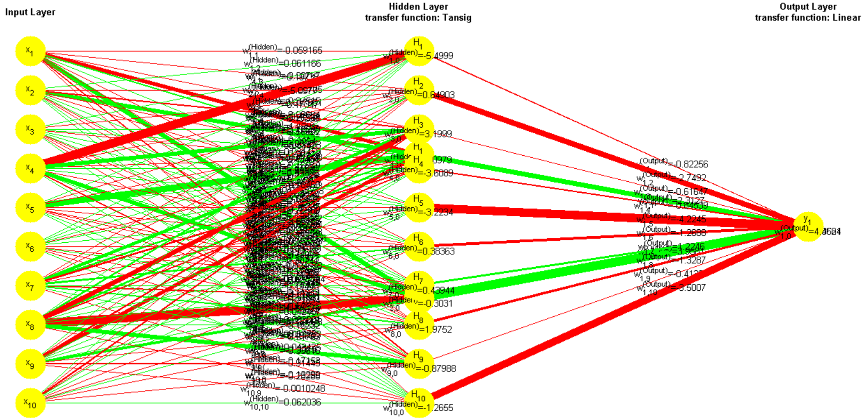
D. $w_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, w_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, w_3 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

E. Don’t know.

Interpreting neural networks can be difficult


Input Layer
Hidden Layer
transfer function: Tansig
Output Layer
transfer function: Linear
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
x_9
x_{10}
H_1
H_2
y_1
$w^{(Hidden)}_{1,1}$=-0.26588
$w^{(Hidden)}_{1,2}$=0.31026
$w^{(Hidden)}_{1,3}$=0.053998
$w^{(Hidden)}_{1,4}$=-0.05456
$w^{(Hidden)}_{1,5}$=0.41148
$w^{(Hidden)}_{1,6}$=-0.073143
$w^{(Hidden)}_{1,7}$=0.17407
$w^{(Hidden)}_{1,8}$=-0.28851
$w^{(Hidden)}_{2,1}$=0.47711
$w^{(Hidden)}_{1,9}$=-0.25637
$w^{(Hidden)}_{2,2}$=-0.022994
$w^{(Hidden)}_{1,10}$=-1.6074e-06
$w^{(Hidden)}_{2,3}$=-0.049888
$w^{(Hidden)}_{2,4}$=0.30677
$w^{(Hidden)}_{2,5}$=-0.034678
$w^{(Hidden)}_{2,6}$=0.025577
$w^{(Hidden)}_{2,7}$=-0.11314
$w^{(Hidden)}_{2,8}$=-0.60574
$w^{(Hidden)}_{2,9}$=0.22443
$w^{(Hidden)}_{2,10}$=0.083263
$w^{(Hidden)}_{1,0}$=1.0979
$w^{(Hidden)}_{2,0}$=-0.3031
$w^{(Output)}_{1,1}$=2.3127
$w^{(Output)}_{1,2}$=3.9581
$w^{(Output)}_{1,0}$=9.8584

Interpreting neural networks can be difficult


Input Layer
Hidden Layer
transfer function: Tansig
Output Layer
transfer function: Linear
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
x_9
x_{10}
H_1
H_2
H_3
H_4
H_5
H_6
H_7
H_8
H_9
H_{10}
y_1
$w^{(Hidden)}_{1,1}$=-0.059165
$w^{(Hidden)}_{1,2}$=0.061166
$w^{(Hidden)}_{10,9}$=0.0010248
$w^{(Hidden)}_{10,10}$=0.062036
$w^{(Hidden)}_{1,0}$=-5.4999
$w^{(Hidden)}_{2,0}$=0.64903
$w^{(Hidden)}_{3,0}$=3.1999
$w^{(Hidden)}_{4,0}$=-3.6089
$w^{(Hidden)}_{5,0}$=-3.2234
$w^{(Hidden)}_{6,0}$=0.38363
$w^{(Hidden)}_{7,0}$=0.43944
$w^{(Hidden)}_{8,0}$=1.9752
$w^{(Hidden)}_{9,0}$=-0.87988
$w^{(Hidden)}_{10,0}$=-1.2655
$w^{(Output)}_{1,1}$=-0.82256
$w^{(Output)}_{1,2}$=-2.7492
$w^{(Output)}_{1,3}$=-0.61647
$w^{(Output)}_{1,5}$=-4.2245
$w^{(Output)}_{1,6}$=-1.2888
$w^{(Output)}_{1,8}$=-1.3287
$w^{(Output)}_{1,10}$=-3.5007

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/)

Midterm practice test

Look at the test on DTU Learn. Note the test is not part of your evaluation.

Midterm question 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.

Midterm question 2

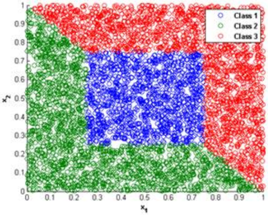
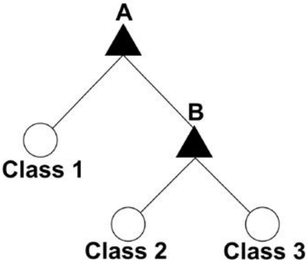


Figure 1



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

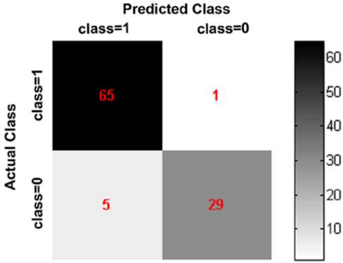
C. A: $\|\boldsymbol{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq 0.25$, B : $\|\boldsymbol{x}_n\|_\infty \leq 1$

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.

Midterm question 3

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



		Predicted Class	
		class=1	class=0
Actual Class	class=1	65	1
	class=0	5	29

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.

Midterm question 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.

Midterm question 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.

Midterm question 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.

Midterm question 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.

Midterm question 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.

Midterm question 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. $y = 1$

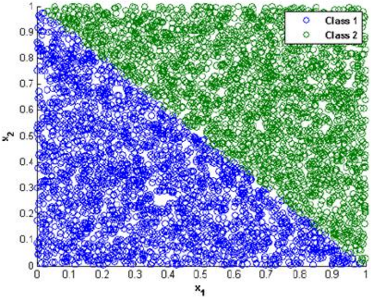
B. $y = 6$

C. $y = 7$

D. $y = 11$

E. Don't know.

Midterm question 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.