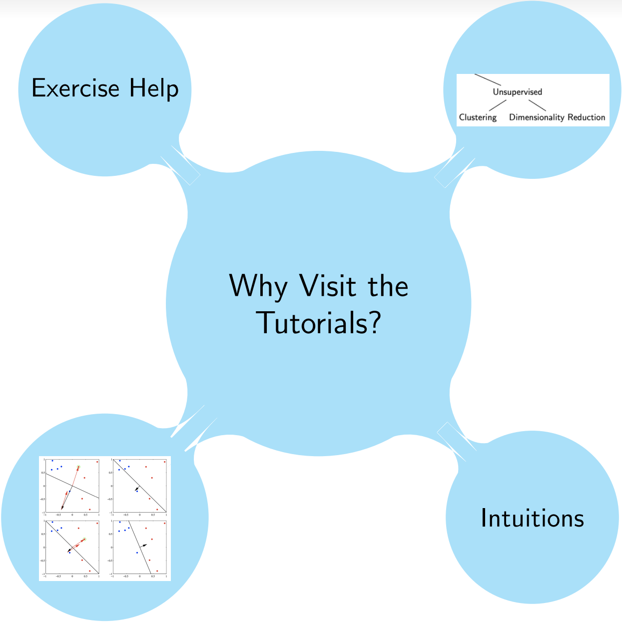
Tutorial Session 1: Linear Classification

Ken, Joanina, Augustin

Organizational Remarks

- **Lecture**:
  Dr. Johannes Niediek (cognitivealgorithms@ml.tu-berlin.de)
- **Tutorials and Gradings**:
  Ken (ken.schreiber@campus.tu-berlin.de)
  Joanina (j.oltersdorff@campus.tu-berlin.de)
  Augustin (augustin.krause@campus.tu-berlin.de)
- We see each other every two weeks!

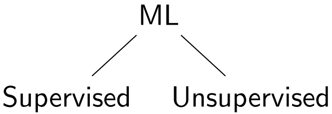

Exercise Help
Unsupervised
Clustering
Dimensionality Reduction
Why Visit the Tutorials?
Intuitions

What is an 'ML-Model'?

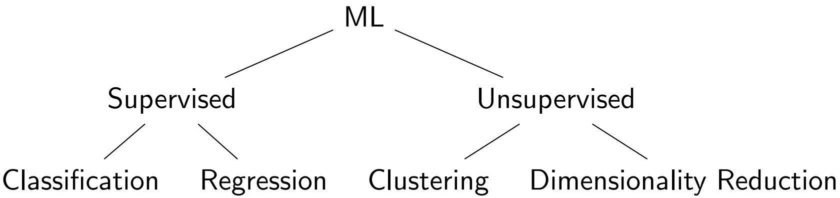
The Tree of CA

ML

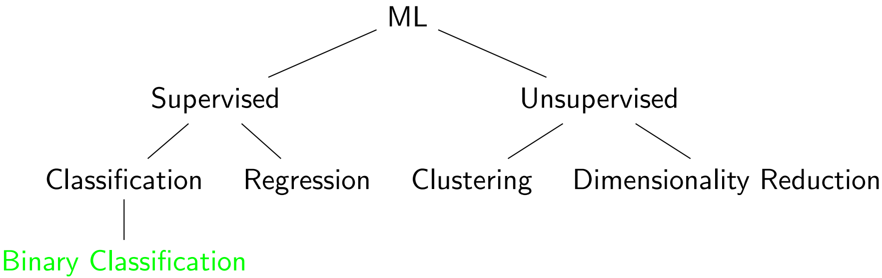
The Tree of CA



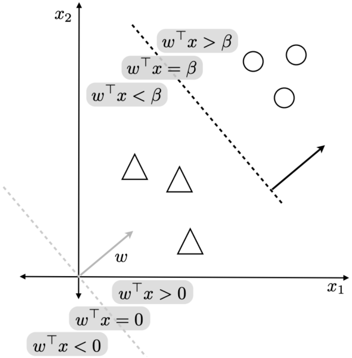
The Tree of CA



The Tree of CA



Interpretation: The Decision Boundary



$\mathbf{x}, \mathbf{w} \in \mathbb{R}^D, \beta \in \mathbb{R}$

$$\mathbf{w}^\top \mathbf{x} - \beta = \begin{cases} > 0 & \text{if } \mathbf{x} \text{ belongs to o} \\ < 0 & \text{if } \mathbf{x} \text{ belongs to } \Delta \end{cases}$$

Points on the decision boundary satisfy $y(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} - \beta = 0$

The Tree of CA

- ML
  - Supervised
    - Classification
      - Binary Classification
        - NCC
    - Regression
  - Unsupervised
    - Clustering
    - Dimensionality Reduction

Task 1 - Example Prototype Classifier

Consider the following data points:

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$\mathbf{x}_1$ and $\mathbf{x}_2$ belong to class -1, while $\mathbf{x}_3$ and $\mathbf{x}_4$ belong to class $+1$.

1. Compute the class means $\mathbf{w}_{-1}$ and $\mathbf{w}_{+1}$.

Task 1 - Example Prototype Classifier

Consider the following data points:

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$\mathbf{x}_1$ and $\mathbf{x}_2$ belong to class -1, while $\mathbf{x}_3$ and $\mathbf{x}_4$ belong to class $+1$.

1. Compute the class means $\mathbf{w}_{-1}$ and $\mathbf{w}_{+1}$.
2. Compute the classification boundary $\mathbf{w}^\top \mathbf{x} - \beta = 0$ of the prototype classifier.

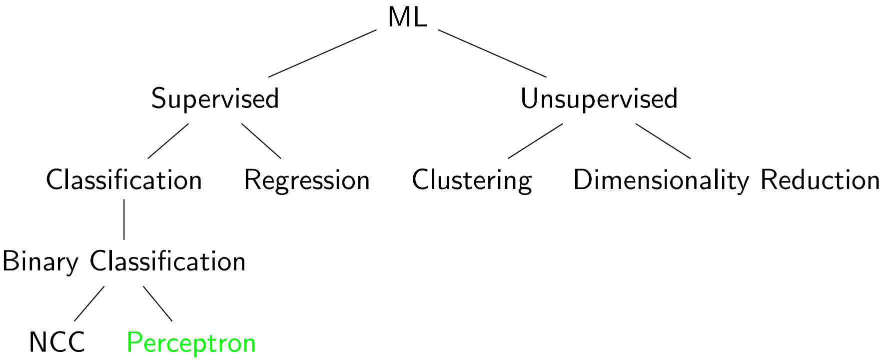
Task 1 - Example Prototype Classifier

3. For each point, compute the assigned class label $\mathrm{sign}(\mathbf{w}^\top \mathbf{x} - \beta)$. Are all points classified correctly?

Task 1 - Example Prototype Classifier

3. For each point, compute the assigned class label $\text{sign}(\mathbf{w}^\top \mathbf{x} - \beta)$. Are all points classified correctly?
4. Sketch the data points, their class means $\mathbf{w}_{-1}$ and $\mathbf{w}_{+1}$, the normal vector $\mathbf{w}$, and the classification boundary.

The Tree of CA



Stochastic Gradient Descent

We do not need to compute the error w.r.t to all data points to do an update. Instead we can choose one data point randomly.

Stochastic Gradient Descent

We do not need to compute the error w.r.t to all data points to do an update. Instead we can choose one data point randomly.
Perceptron SGD:

1. Initialize $\mathbf{w}^{\text{old}}$ (randomly, $1/n$, ...)

Stochastic Gradient Descent

We do not need to compute the error w.r.t to all data points to do an update. Instead we can choose one data point randomly.
Perceptron SGD:

1. Initialize $\mathbf{w}^{\text{old}}$ (randomly, $1/n$, ...)
2. While there are misclassified data points (or until stopping criterion is reached)

   Pick a random misclassified data point $\mathbf{x}_m$

   Descent in direction of the gradient at single data point $\mathbf{x}_m$

Stochastic Gradient Descent

We do not need to compute the error w.r.t to all data points to do an update. Instead we can choose one data point randomly. Perceptron SGD:

1. Initialize $\mathbf{w}^{\text{old}}$ (randomly, $1/n$, ...)
2. While there are misclassified data points (or until stopping criterion is reached)

   Pick a random misclassified data point $\mathbf{x}_m$
   Descent in direction of the gradient at single data point $\mathbf{x}_m$

$$
\begin{aligned}
\mathcal{E}_m(\mathbf{w}) &= -\mathbf{w}^\top \mathbf{x}_m y_m \\
\nabla\mathcal{E}_m(\mathbf{w}) &= -\mathbf{x}_m y_m \\
\mathbf{w}^{\text{new}} &\leftarrow \mathbf{w}^{\text{old}} - \eta\nabla\mathcal{E}_m(\mathbf{w}^{\text{old}}) = \mathbf{w}^{\text{old}} + \eta\mathbf{x}_m y_m
\end{aligned}
$$

Task 3 - Convergence of the Perceptron

1. We denote a hyperplane by $\mathbf{w}^\top \mathbf{x} = 0$. Show that there exists a $\mathbf{w}_{\text{sep}}$ such that:

$$\mathbf{w}_{\text{sep}}^\top \mathbf{x}_i y_i \geq \|\mathbf{x}_i\|^2, \forall i \in \{1, \ldots, N\}.$$

Task 3 - Convergence of the Perceptron

1. We denote a hyperplane by $\mathbf{w}^\top \mathbf{x} = 0$. Show that there exists a $\mathbf{w}_{\text{sep}}$ such that:

$$\mathbf{w}_{\text{sep}}^\top \mathbf{x}_i y_i \geq \|\mathbf{x}_i\|^2, \forall i \in \{1, \dots, N\}.$$

2. Given a current $\mathbf{w}_{\text{old}} \in \mathbb{R}$, the perceptron algorithm identifies a point $\mathbf{x}_m$ that is misclassified, and produces the update rule $\mathbf{w}_{\text{new}} = \mathbf{w}_{\text{old}} + \eta \mathbf{x}_m y_m$. Using this update rule, show that

$$\|\mathbf{w}_{\text{new}} - \mathbf{w}_{\text{sep}}\|^2 \leq \|\mathbf{w}_{\text{old}} - \mathbf{w}_{\text{sep}}\|^2 - \|\mathbf{x}_m\|^2 \tag{1}$$

This implies that the perceptron algorithm converges to a separating hyperplane in a finite number of steps.

Task 2 - The linear classification boundary

Consider a linear classification boundary $\mathbf{w}^\top \mathbf{x} - \beta = 0$. Draw a sketch in 2D to visualize the classification boundary and answer the following questions:

1. Suppose $\beta = 0$ and $\|\mathbf{w}\| = 1$. How large is the distance of a point $\mathbf{z}$ to the classification boundary?
2. How large is the distance of a point $\mathbf{z}$ to the classification boundary if $\|\mathbf{w}\| = 1$ but $\beta \neq 0$?
3. How large is the distance of a point $\mathbf{z}$ to the classification boundary for arbitrary β and $\mathbf{w}$?
4. How large is the distance between a classification boundary $\mathbf{w}$ and the origin for arbitrary β and $\mathbf{w}$?
5. Is β in the general case the intercept of the classification boundary $\mathbf{w}^\top \mathbf{x} - \beta = 0$ with the x_2-axis? If yes, explain why. If not, give a counter-example.

Task 2 - The linear classification boundary

1. $\beta = 0$ and $\|\mathbf{w}\| = 1$

Task 2 - The linear classification boundary

1. $\beta = 0$ and $\|\mathbf{w}\| = 1$
2. $\beta \neq 0$ and $\|\mathbf{w}\| = 1$

Task 2 - The linear classification boundary

1. $\beta = 0$ and $\|\mathbf{w}\| = 1$
2. $\beta \neq 0$ and $\|\mathbf{w}\| = 1$
3. arbitrary β and $\mathbf{w}$

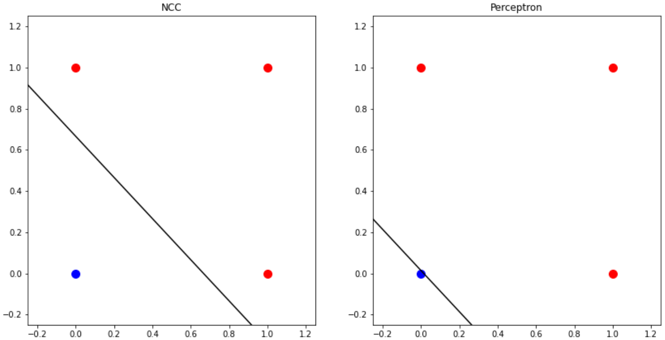
Task 2 - The linear classification boundary

1. $\beta = 0$ and $\|\mathbf{w}\| = 1$
2. $\beta \neq 0$ and $\|\mathbf{w}\| = 1$
3. arbitrary β and $\mathbf{w}$
4. distance between boundary and the origin for arbitrary β and $\mathbf{w}$

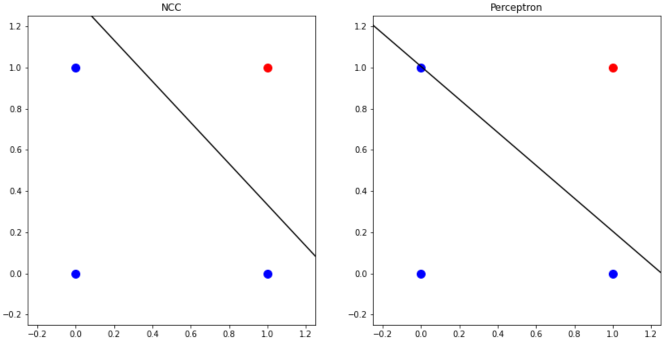
Task 2 - The linear classification boundary

1. $\beta = 0$ and $\|\mathbf{w}\| = 1$
2. $\beta \neq 0$ and $\|\mathbf{w}\| = 1$
3. arbitrary β and $\mathbf{w}$
4. distance between boundary and the origin for arbitrary β and $\mathbf{w}$
5. Is β intercept of classification boundary with the x_2-axis?

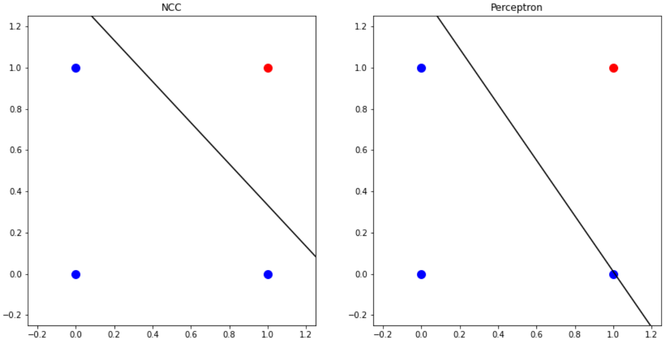
NCC and Perceptron: OR



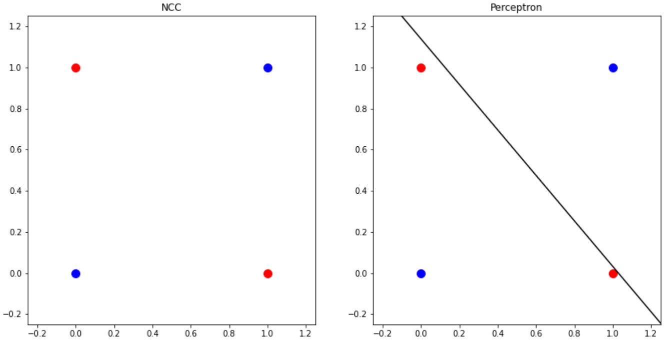
NCC and Perceptron: AND (1)



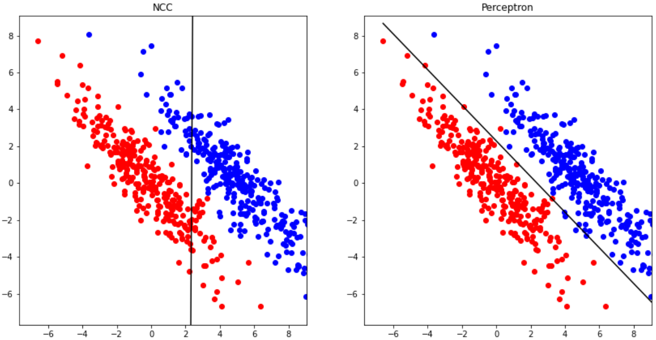
NCC and Perceptron: AND (2)



NCC and Perceptron: XOR



NCC and Perceptron



NCC and Perceptron

	NCC	**Perceptron**
Problem	Classification	Classification(, Regression)
Model	$y = \text{sign}(\mathbf{w}^T\mathbf{x})$	$y = f(\mathbf{w}^T\mathbf{x})$
Error	distance to $\mathbf{w}_{+1}, \mathbf{w}_{-1}$	$-\sum_{m\in M}\mathbf{w}^T\mathbf{x}_m y_m$
Optimization	closed form	SGD
Result	always the same	can differ
Application	Cancer Prediction[1]	NLP[2]

[1] (Tibshirani et al., 2002)
[2] (Collins, 2002)

References

Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics, 2002.

Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10): 6567–6572, 2002.