Homework 4 Recap & Decision Theory

Nemanja Bartolovic

06.05.2020

Overview

- Homework 4 Recap
  - Problem set 2 (Questions 4-7)
  - Problem set 5 (Questions 16-19)

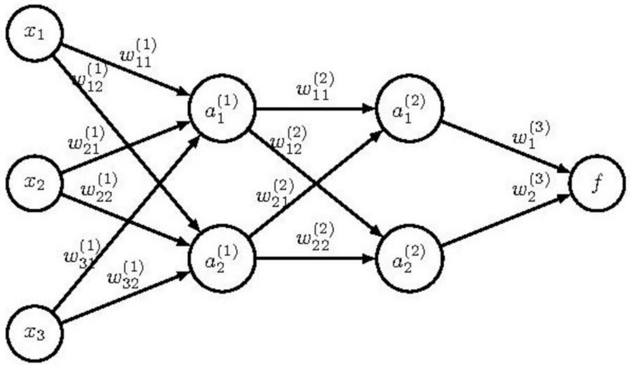
- Decision Theory
  - Theory recap
  - Examples

Overview

- Homework 4 Recap
  - **Problem set 2 (Questions 4-7)**
  - Problem set 5 (Questions 16-19)
- Decision Theory
  - Theory recap
  - Examples

Homework 4 Recap: Question 4

After some semesters of attending exams, Xiaoming finds out he can not collect enough training samples. So he decides to use a dropout technique to reduce overfitting of his model. In particular, he applies dropout for the 2nd hidden layer ($a_1^{(2)}$and $a_2^{(2)}$) with the probability of the corresponding neuron being retained being 0.4. Help him compute the expected value of the loss in this case, given training example x_1, x_2, x_3 and grade y, by answering the following questions.



Homework 4 Recap: Question 4

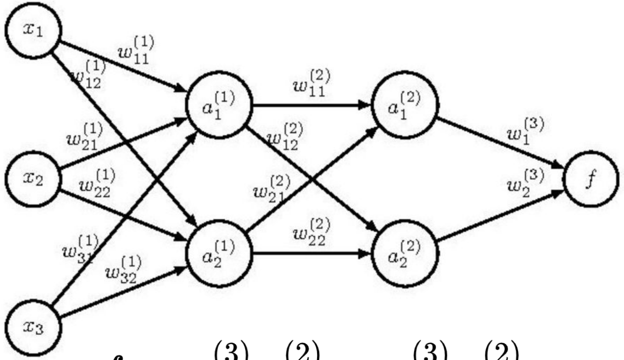
$$\mathbb{E}_{a_1^{(2)}, a_2^{(2)}}(f) = ?$$

Homework 4 Recap: Question 4

$$\mathbb{E}_{a_1^{(2)},a_2^{(2)}}(f) = ?$$

- If a neuron is being retained with p = 0.4, then it will be dropped with probability 1 - p = 0.6
- Remember the linearity of expected value

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$
$$\mathbb{E}[aX] = a\mathbb{E}[X], a \in \mathbb{R}$$
$$\mathbb{E}[a] = a, a \in \mathbb{R}$$



$$f = w_1^{(3)}\alpha_1^{(2)} + w_2^{(3)}\alpha_2^{(2)}$$

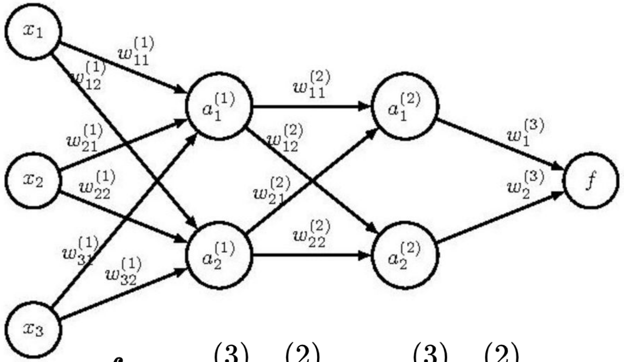
Homework 4 Recap: Question 4

$$\mathbb{E}_{a_1^{(2)},a_2^{(2)}}(f) = ?$$

- If a neuron is being retained with p = 0.4, then it will be dropped with probability 1 - p = 0.6
- Remember the linearity of expected value

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$
$$\mathbb{E}[aX] = a\mathbb{E}[X], a \in \mathbb{R}$$
$$\mathbb{E}[a] = a, a \in \mathbb{R}$$

$$\mathbb{E}(f) = \mathbb{E}(w_1^{(3)}\alpha_1^{(2)}) + \mathbb{E}(w_2^{(3)}\alpha_2^{(2)})$$



$$f = w_1^{(3)}\alpha_1^{(2)} + w_2^{(3)}\alpha_2^{(2)}$$

Homework 4 Recap: Question 4

- Now we can directly calculate the expected values (comparable to Bernoulli distribution)

$$\mathbb{E}(w_1^{(3)}\alpha_1^{(2)}) = p \cdot w_1^{(3)}\alpha_1^{(2)} + (1-p)\cdot 0 = pw_1^{(3)}\alpha_1^{(2)}$$
$$\mathbb{E}(w_2^{(3)}\alpha_2^{(2)}) = p \cdot w_2^{(3)}\alpha_2^{(2)} + (1-p)\cdot 0 = pw_2^{(3)}\alpha_2^{(2)}$$

Homework 4 Recap: Question 4

- Now we can directly calculate the expected values (comparable to Bernoulli distribution)

$$\mathbb{E}(w_1^{(3)}\alpha_1^{(2)}) = p \cdot w_1^{(3)}\alpha_1^{(2)} + (1-p)\cdot 0 = pw_1^{(3)}\alpha_1^{(2)}$$
$$\mathbb{E}(w_2^{(3)}\alpha_2^{(2)}) = p \cdot w_2^{(3)}\alpha_2^{(2)} + (1-p)\cdot 0 = pw_2^{(3)}\alpha_2^{(2)}$$

- Finally, just plug everything back in

$$\begin{aligned}\mathbb{E}(f) &= \mathbb{E}(w_1^{(3)}\alpha_1^{(2)}) + \mathbb{E}(w_2^{(3)}\alpha_2^{(2)}) \\ &= p(w_1^{(3)}\alpha_1^{(2)} + w_2^{(3)}\alpha_2^{(2)}) \\ &= 0.4(w_1^{(3)}\alpha_1^{(2)} + w_2^{(3)}\alpha_2^{(2)})\end{aligned}$$

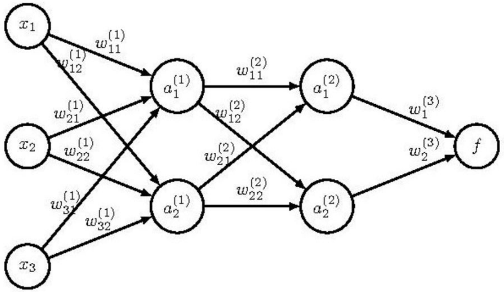
Homework 4 Recap: Question 5

$$\mathrm{Var}_{a_1^{(2)}, a_2^{(2)}}(f) = ?$$

Homework 4 Recap: Question 5

$$\mathrm{Var}_{a_1^{(2)}, a_2^{(2)}}(f) = ?$$

- Remember the definition of variance and its properties



$$
\begin{aligned}
\mathrm{Var}[X+Y] &= \mathrm{Var}[X] + \mathrm{Var}[Y] \\
\mathrm{Var}[aX] &= a^2 \mathrm{Var}[X], a \in \mathbb{R} \\
\mathrm{Var}[X] &= \mathbb{E}[(X-\mu)^2], \qquad \mu = \mathbb{E}[X] \\
\Rightarrow \mathrm{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[2X \cdot \mathbb{E}[X]] + \mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2
\end{aligned}
$$

$$f = w_1^{(3)} \alpha_1^{(2)} + w_2^{(3)} \alpha_2^{(2)}$$

$$\mathrm{Var}(f) = \mathrm{Var}(w_1^{(3)} \alpha_1^{(2)}) + \mathrm{Var}(w_2^{(3)} \alpha_2^{(2)})$$

Homework 4 Recap: Question 5

- We can directly calculate the variance using expected values from Question 4

$$\text{Var}(w_1^{(3)}\alpha_1^{(2)}) = \mathbb{E}[(w_1^{(3)}\alpha_1^{(2)})^2] - \mathbb{E}[w_1^{(3)}\alpha_1^{(2)}]^2$$

$$\mathbb{E}(w_1^{(3)}\alpha_1^{(2)}) = p \cdot w_1^{(3)}\alpha_1^{(2)} + (1-p) \cdot 0 = pw_1^{(3)}\alpha_1^{(2)}$$
$$\mathbb{E}(w_2^{(3)}\alpha_2^{(2)}) = p \cdot w_2^{(3)}\alpha_2^{(2)} + (1-p) \cdot 0 = pw_2^{(3)}\alpha_2^{(2)}$$

Homework 4 Recap: Question 5

- We can directly calculate the variance using expected values from Question 4

$$\mathrm{Var}(w_1^{(3)}\alpha_1^{(2)}) = \mathbb{E}[(w_1^{(3)}\alpha_1^{(2)})^2] - \mathbb{E}[w_1^{(3)}\alpha_1^{(2)}]^2$$
$$= p(w_1^{(3)}\alpha_1^{(2)})^2 - p^2(w_1^{(3)}\alpha_1^{(2)})^2$$

$$\mathbb{E}(w_1^{(3)}\alpha_1^{(2)}) = p \cdot w_1^{(3)}\alpha_1^{(2)} + (1-p)\cdot 0 = pw_1^{(3)}\alpha_1^{(2)}$$
$$\mathbb{E}(w_2^{(3)}\alpha_2^{(2)}) = p \cdot w_2^{(3)}\alpha_2^{(2)} + (1-p)\cdot 0 = pw_2^{(3)}\alpha_2^{(2)}$$

Homework 4 Recap: Question 5

- We can directly calculate the variance using expected values from Question 4

$$\begin{aligned}\mathrm{Var}(w_1^{(3)}\alpha_1^{(2)}) &= \mathbb{E}[(w_1^{(3)}\alpha_1^{(2)})^2] - \mathbb{E}[w_1^{(3)}\alpha_1^{(2)}]^2 \\ &= p(w_1^{(3)}\alpha_1^{(2)})^2 - p^2(w_1^{(3)}\alpha_1^{(2)})^2 \\ &= p(1-p)(w_1^{(3)}\alpha_1^{(2)})^2\end{aligned}$$

$$\mathbb{E}(w_1^{(3)}\alpha_1^{(2)}) = p \cdot w_1^{(3)}\alpha_1^{(2)} + (1-p)\cdot 0 = pw_1^{(3)}\alpha_1^{(2)}$$
$$\mathbb{E}(w_2^{(3)}\alpha_2^{(2)}) = p \cdot w_2^{(3)}\alpha_2^{(2)} + (1-p)\cdot 0 = pw_2^{(3)}\alpha_2^{(2)}$$

Homework 4 Recap: Question 5

- We can directly calculate the variance using expected values from Question 4

$$\begin{aligned}\mathrm{Var}(w_1^{(3)}\alpha_1^{(2)}) &= \mathbb{E}[(w_1^{(3)}\alpha_1^{(2)})^2] - \mathbb{E}[w_1^{(3)}\alpha_1^{(2)}]^2 \\ &= p(w_1^{(3)}\alpha_1^{(2)})^2 - p^2(w_1^{(3)}\alpha_1^{(2)})^2 \\ &= p(1-p)(w_1^{(3)}\alpha_1^{(2)})^2\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(w_2^{(3)}\alpha_2^{(2)}) &= \mathbb{E}[(w_2^{(3)}\alpha_2^{(2)})^2] - \mathbb{E}[w_2^{(3)}\alpha_2^{(2)}]^2 \\ &= p(w_2^{(3)}\alpha_2^{(2)})^2 - p^2(w_2^{(3)}\alpha_2^{(2)})^2 \\ &= p(1-p)(w_2^{(3)}\alpha_2^{(2)})^2\end{aligned}$$

Homework 4 Recap: Question 5

- We can directly calculate the variance using expected values from Question 4

$$\begin{aligned}\mathrm{Var}(w_1^{(3)}\alpha_1^{(2)}) &= \mathbb{E}[(w_1^{(3)}\alpha_1^{(2)})^2] - \mathbb{E}[w_1^{(3)}\alpha_1^{(2)}]^2 \\ &= p(w_1^{(3)}\alpha_1^{(2)})^2 - p^2(w_1^{(3)}\alpha_1^{(2)})^2 \\ &= p(1-p)(w_1^{(3)}\alpha_1^{(2)})^2\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(w_2^{(3)}\alpha_2^{(2)}) &= \mathbb{E}[(w_2^{(3)}\alpha_2^{(2)})^2] - \mathbb{E}[w_2^{(3)}\alpha_2^{(2)}]^2 \\ &= p(w_2^{(3)}\alpha_2^{(2)})^2 - p^2(w_2^{(3)}\alpha_2^{(2)})^2 \\ &= p(1-p)(w_2^{(3)}\alpha_2^{(2)})^2\end{aligned}$$

- Finally, just plug everything back in

$$\mathrm{Var}(f) = \mathrm{Var}(w_1^{(3)}\alpha_1^{(2)}) + \mathrm{Var}(w_2^{(3)}\alpha_2^{(2)})$$

Homework 4 Recap: Question 5

- We can directly calculate the variance using expected values from Question 4

$$\begin{aligned}\mathrm{Var}(w_1^{(3)}\alpha_1^{(2)}) &= \mathbb{E}[(w_1^{(3)}\alpha_1^{(2)})^2] - \mathbb{E}[w_1^{(3)}\alpha_1^{(2)}]^2 \\ &= p(w_1^{(3)}\alpha_1^{(2)})^2 - p^2(w_1^{(3)}\alpha_1^{(2)})^2 \\ &= p(1-p)(w_1^{(3)}\alpha_1^{(2)})^2\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(w_2^{(3)}\alpha_2^{(2)}) &= \mathbb{E}[(w_2^{(3)}\alpha_2^{(2)})^2] - \mathbb{E}[w_2^{(3)}\alpha_2^{(2)}]^2 \\ &= p(w_2^{(3)}\alpha_2^{(2)})^2 - p^2(w_2^{(3)}\alpha_2^{(2)})^2 \\ &= p(1-p)(w_2^{(3)}\alpha_2^{(2)})^2\end{aligned}$$

- Finally, just plug everything back in

$$\begin{aligned}\mathrm{Var}(f) &= \mathrm{Var}(w_1^{(3)}\alpha_1^{(2)}) + \mathrm{Var}(w_2^{(3)}\alpha_2^{(2)}) \\ &= p(1-p)((w_1^{(3)}\alpha_1^{(2)})^2 + (w_2^{(3)}\alpha_2^{(2)})^2) \\ &= 0.24((w_1^{(3)}\alpha_1^{(2)})^2 + (w_2^{(3)}\alpha_2^{(2)})^2)\end{aligned}$$

Homework 4 Recap: Question 6

$$\mathbb{E}(L) = ?$$

- Solution 1: we use the properties of expected value and variance that we know

$$\begin{aligned} L &= (Y - f)^2 \\ &= Y^2 - 2Yf + f^2 \end{aligned}$$

Homework 4 Recap: Question 6

$$\mathbb{E}(L) = ?$$

- Solution 1: we use the properties of expected value and variance that we know

$$\begin{aligned} L &= (Y - f)^2 \\ &= Y^2 - 2Yf + f^2 \end{aligned}$$

$$\Rightarrow$$

$$\begin{aligned} \mathbb{E}(L) &= \mathbb{E}(Y^2) - \mathbb{E}(2Yf) + \mathbb{E}(f^2) \\ &= Y^2 - 2Y\mathbb{E}(f) + \mathbb{E}(f^2) \end{aligned}$$

Homework 4 Recap: Question 6

$$\mathbb{E}(L) = ?$$

- Solution 1: we use the properties of expected value and variance that we know

$$\begin{aligned} L &= (Y - f)^2 \\ &= Y^2 - 2Yf + f^2 \\ &\Rightarrow \\ \mathbb{E}(L) &= \mathbb{E}(Y^2) - \mathbb{E}(2Yf) + \mathbb{E}(f^2) \\ &= Y^2 - 2Y\mathbb{E}(f) + \mathbb{E}(f^2) \end{aligned}$$

- Solution 2: Combine the variance and expected value identity and Solution 1

$$\begin{aligned} \mathrm{Var}(f) &= \mathbb{E}(f^2) - \mathbb{E}(f)^2 \\ \mathbb{E}(f^2) &= \mathrm{Var}(f) + \mathbb{E}(f)^2 \\ &\Rightarrow \\ \mathbb{E}(L) &= \mathbb{E}(Y^2) - \mathbb{E}(2Yf) + \mathbb{E}(f^2) \\ &= Y^2 - 2Y\mathbb{E}(f) + \mathrm{Var}(f) + \mathbb{E}(f)^2 \end{aligned}$$

- Tick both correct answers!

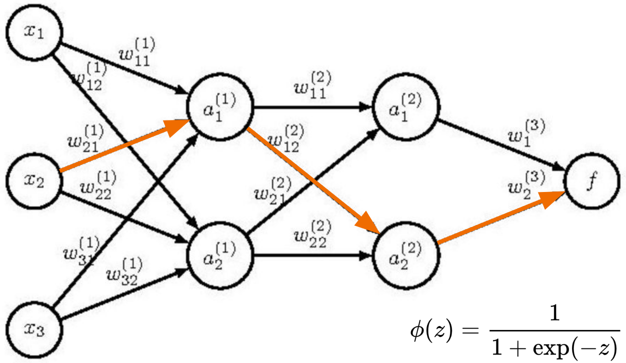
Homework 4 Recap: Question 7

- Help Xiaoming compute the derivative

$$\frac{dL}{dw_{21}^{(1)}} = ?$$

- We will use the chain derivation rule

$$\frac{df(g(x))}{dx} = \left.\frac{df}{dy}\right|_{g(x)} \cdot \left.\frac{dy}{dx}\right|_{x}^{g(x)} = f'(g(x))g'(x)$$



$$\phi(z) = \frac{1}{1+\exp(-z)}$$

Homework 4 Recap: Question 7

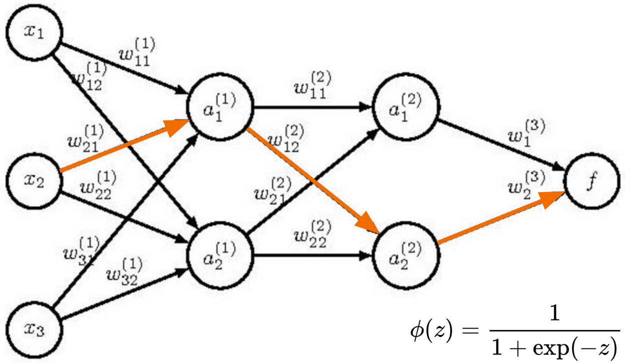
- Help Xiaoming compute the derivative

$$\frac{dL}{dw_{21}^{(1)}} = ?$$

$$\frac{dL}{dw_{21}^{(1)}} = 2(f-y)\cdot\frac{df}{dw_{21}^{(1)}}$$
$$= 2(f-y)\cdot\frac{d(w_2^{(3)}\alpha_2^{(2)})}{dw_{21}^{(1)}}$$

- We will use the chain derivation rule

$$\frac{df(g(x))}{dx} = \left.\frac{df}{dy}\right|_{g(x)} \cdot \left.\frac{dy}{dx}\right|_{x}^{g(x)} = f'(g(x))g'(x)$$



$$\phi(z) = \frac{1}{1+\exp(-z)}$$

Homework 4 Recap: Question 7

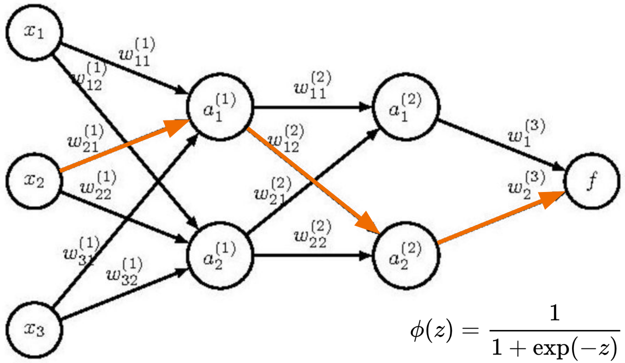
- Help Xiaoming compute the derivative

$$\frac{dL}{dw_{21}^{(1)}} = ?$$

$$\begin{aligned}\frac{dL}{dw_{21}^{(1)}} &= 2(f-y)\cdot\frac{df}{dw_{21}^{(1)}}\\ &= 2(f-y)\cdot\frac{d(w_2^{(3)}\alpha_2^{(2)})}{dw_{21}^{(1)}}\\ &= 2(f-y)w_2^{(3)}\phi'(w_{12}^{(2)}\alpha_1^{(1)}+w_{22}^{(2)}\alpha_2^{(1)})\frac{d(w_{12}^{(2)}\alpha_1^{(1)})}{dw_{21}^{(1)}}\end{aligned}$$

- We will use the chain derivation rule

$$\frac{df(g(x))}{dx} = \left.\frac{df}{dy}\right|_{g(x)} \cdot \left.\frac{dy}{dx}\right|_{x}^{g(x)} = f'(g(x))g'(x)$$



$$\phi(z) = \frac{1}{1+\exp(-z)}$$

Homework 4 Recap: Question 7

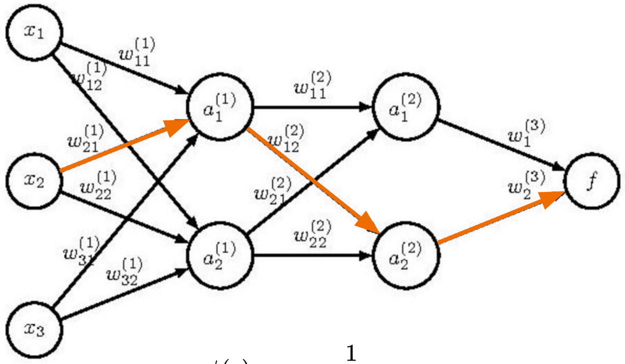
- Help Xiaoming compute the derivative

$$\frac{dL}{dw_{21}^{(1)}} = ?$$

$$\begin{aligned}
\frac{dL}{dw_{21}^{(1)}} &= 2(f-y)\cdot\frac{df}{dw_{21}^{(1)}} \\
&= 2(f-y)\cdot\frac{d(w_2^{(3)}\alpha_2^{(2)})}{dw_{21}^{(1)}} \\
&= 2(f-y)w_2^{(3)}\phi'(w_{12}^{(2)}\alpha_1^{(1)} + w_{22}^{(2)}\alpha_2^{(1)})\frac{d(w_{12}^{(2)}\alpha_1^{(1)})}{dw_{21}^{(1)}} \\
&= 2(f-y)w_2^{(3)}\phi'(w_{12}^{(2)}\alpha_1^{(1)} + w_{22}^{(2)}\alpha_2^{(1)})w_{12}^{(2)}\phi'(w_{11}^{(1)}x_1 + w_{21}^{(1)}x_2 + w_{31}^{(1)}x_3)\frac{d(w_{21}^{(1)}x_2)}{dw_{21}^{(1)}} \\
&= 2(f-y)w_2^{(3)}\phi'(w_{12}^{(2)}\alpha_1^{(1)} + w_{22}^{(2)}\alpha_2^{(1)})w_{12}^{(2)}\phi'(w_{11}^{(1)}x_1 + w_{21}^{(1)}x_2 + w_{31}^{(1)}x_3)x_2
\end{aligned}$$

- We will use the chain derivation rule

$$\frac{df(g(x))}{dx} = \left.\frac{df}{dy}\right|_{g(x)} \cdot \left.\frac{dy}{dx}\right|_{x}^{g(x)} = f'(g(x))g'(x)$$



$$\phi(z) = \frac{1}{1+\exp(-z)}$$

$$\phi'(z) = \phi(z)(1-\phi(z))$$

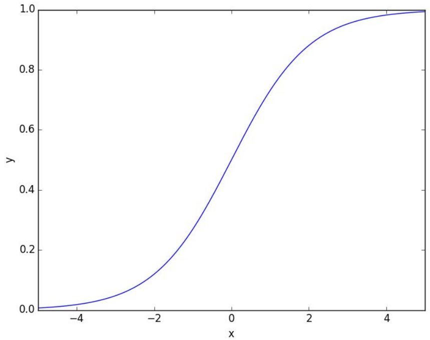
Overview

- Homework 4 Recap
  - Problem set 2 (Questions 4-7)
  - **Problem set 5 (Questions 16-19)**
- Decision Theory
  - Theory recap
  - Examples

Homework 4 Recap: Questions 16-19

- In these questions, we will model boolean operations using neural networks
- We are considering:
  - sigmoid activation function

$$\varphi(z) = \frac{1}{1 + \exp(-z)}$$



  - neuron activations with a bias

$$v_j^l = \varphi(w_0 + \sum_{i \in L_{l-1}} w_{j,i} v_i^{l-1})$$

  - rounding at the end

$$y(z) = \begin{cases} 0 & \varphi(z) < 0.5 \\ 1 & \varphi(z) \geq 0.5 \end{cases} = \text{round}(z)$$

Homework 4 Recap: Question 16

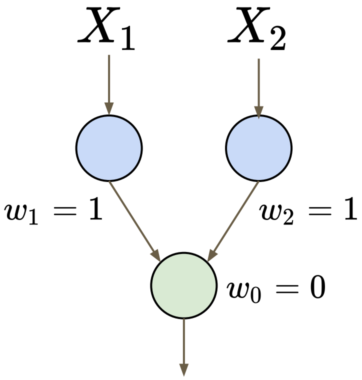
- Designing OR function of two boolean variables $Y = X_1 \vee X_2$
- Allowed weights: -0.5, 0 or 1
- First attempt, directly from the truth table, ignoring bias

X_1	X_2	Y
0	0	0
0	1	1
1	0	1
1	1	1

Homework 4 Recap: Question 16

- Designing OR function of two boolean variables $Y = X_1 \vee X_2$
- Allowed weights: -0.5, 0 or 1
- First attempt, directly from the truth table, ignoring bias

X_1	X_2	Y
0	0	0
0	1	1
1	0	1
1	1	1

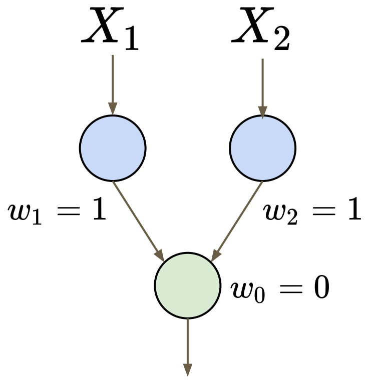


$$y = \text{round}\left(\varphi(0 + 1 \cdot X_1 + 1 \cdot X_2)\right)$$

Homework 4 Recap: Question 16

- Designing OR function of two boolean variables $Y = X_1 \vee X_2$
- Allowed weights: -0.5, 0 or 1
- First attempt, directly from the truth table, ignoring bias

X_1	X_2	Y
0	0	0
0	1	1
1	0	1
1	1	1



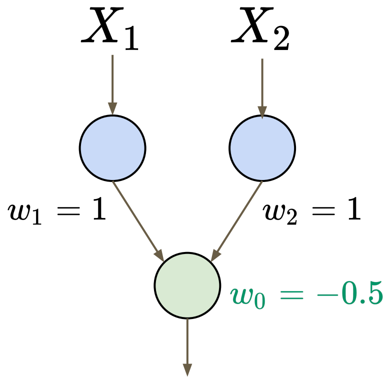
$$y = \text{round}\left(\varphi(0 + 1 \cdot X_1 + 1 \cdot X_2)\right)$$

X_1	X_2	φ	y	
0	0	0.5	1	Wrong!
0	1	≈ 0.73	1	
1	0	≈ 0.73	1	
1	1	≈ 0.88	1	

Homework 4 Recap: Question 16

- Designing OR function of two boolean variables $Y = X_1 \vee X_2$
- Allowed weights: -0.5, 0 or 1
- Use a small negative bias to "push" the sigmoid output below 0.5 when both inputs are zero

X_1	X_2	Y
0	0	0
0	1	1
1	0	1
1	1	1



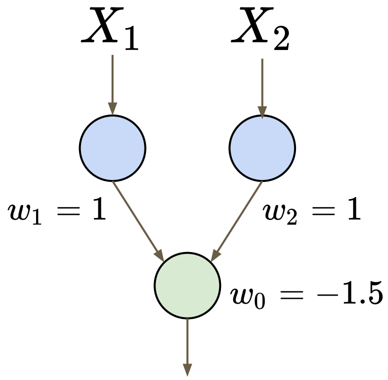
X_1	X_2	φ	y
0	0	≈ 0.38	0
0	1	≈ 0.62	1
1	0	≈ 0.62	1
1	1	≈ 0.82	1

$$y = \text{round}\left(\varphi(-0.5 + 1 \cdot X_1 + 1 \cdot X_2)\right)$$

Homework 4 Recap: Question 17

- Designing AND function of two boolean variables $Y = X_1 \wedge X_2$
- Allowed weights: -2, -1.5, -1, -0.5, 0, 0.5 or 1
- Same approach as before, we need to use a negative bias to move the decision boundary even further (there is more than 1 solution)

X_1	X_2	Y
0	0	0
0	1	0
1	0	0
1	1	1



X_1	X_2	φ	y
0	0	≈ 0.18	0
0	1	≈ 0.38	0
1	0	≈ 0.38	0
1	1	≈ 0.62	1

$$y = \text{round}\left(\varphi(-1.5 + 1 \cdot X_1 + 1 \cdot X_2)\right)$$

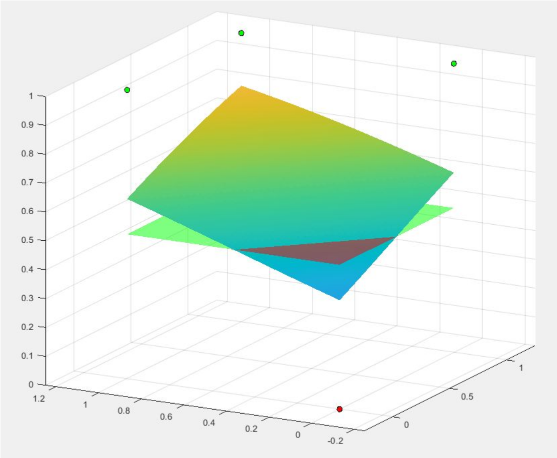
Homework 4 Recap: Question 19

- Designing XOR function of two boolean variables $Y = X_1 \oplus X_2$
- However, XOR is not linearly separable
- An approach with no hidden layers will not work
- So how many hidden layers do we need at least?

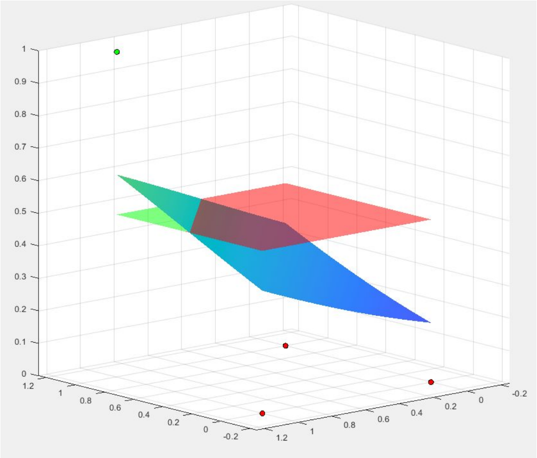
X_1	X_2	Y
0	0	0
0	1	1
1	0	1
1	1	0

Homework 4 Recap: Question 19

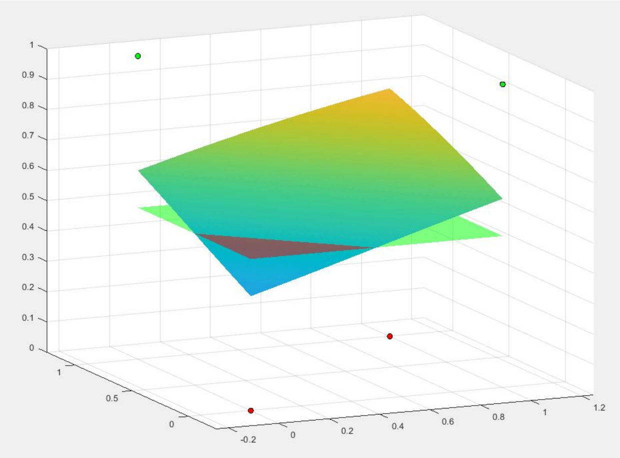
- XOR is not linearly separable



OR function



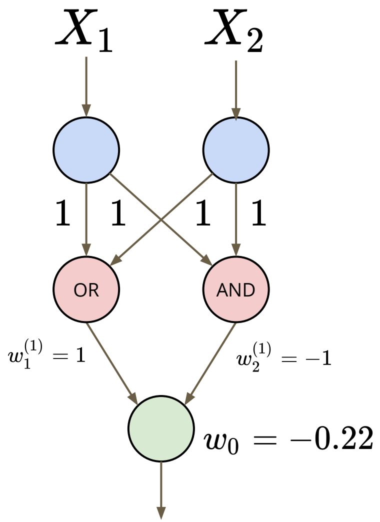
AND function



XOR function

Homework 4 Recap: Question 19

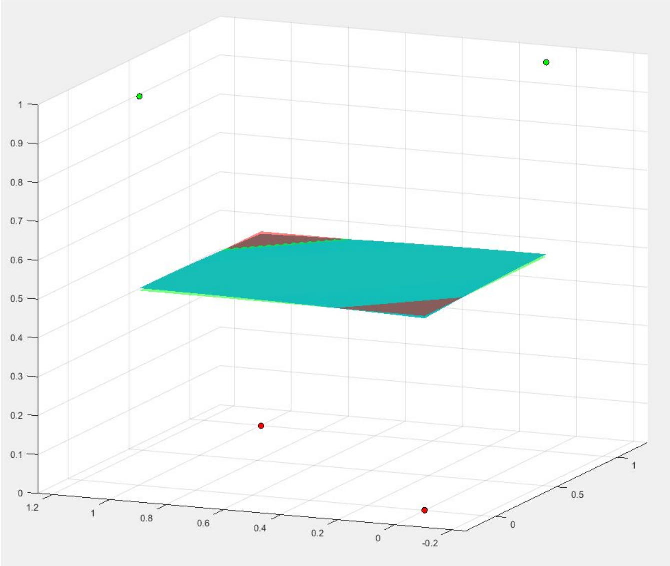
An idea: looking at the truth tables, we could "decompose" XOR as "OR - AND"



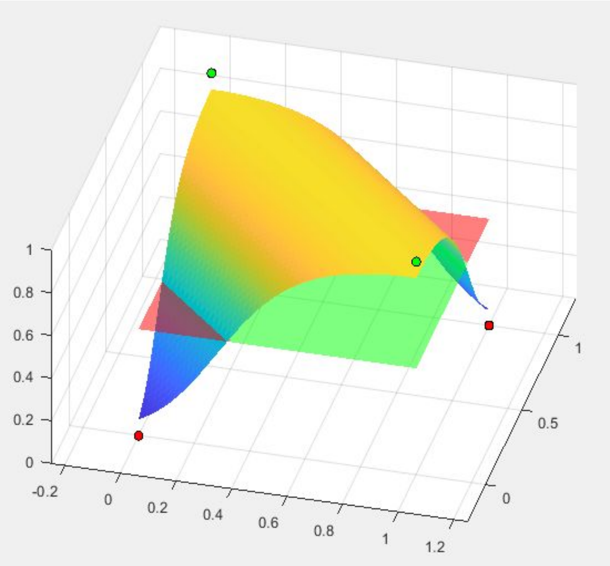
$$y = \text{round}\left(\varphi(-0.22 + 1 \cdot \alpha_1 - 1 \cdot \alpha_2)\right)$$

X_1	X_2	OR	AND	OR − AND	w_0 + OR − AND	φ	y
0	0	≈ 0.38	≈ 0.18	≈ 0.195	≈ -0.025	≈ 0.493	0
0	1	≈ 0.62	≈ 0.38	≈ 0.245	≈ 0.025	≈ 0.507	1
1	0	≈ 0.62	≈ 0.38	≈ 0.245	≈ 0.025	≈ 0.507	1
1	1	≈ 0.82	≈ 0.62	≈ 0.195	≈ -0.025	≈ 0.493	0

Homework 4 Recap: Question 19



our XOR network output



same input, scaled up

Overview

- Homework 4 Recap
  - Problem set 2 (Questions 4-7)
  - Problem set 5 (Questions 16-19)
- Decision Theory
  - **Theory recap**
  - Examples

Decision Theory

- Reason about risk and **decisions under uncertainty**
- After we have estimated a probabilistic model of the data through supervised learning, it would be useful to see how “risky” it would be to use it for making decisions
- Therefore, we are interested in establishing a way to **estimate the risk or the cost** of using our model in uncertain conditions
- Quantifies trade-offs between probabilistic classifications and the costs of decisions that are derived from them

Decision Theory

- We start with the following
  - Estimated conditional probabilities for a set of labels $P(y|X), \quad y \in \mathcal{Y}$
  - Set of actions that we can take $\mathcal{A}$
    - not necessarily equal to $\mathcal{Y}$
  - Associated risk/cost for taking these actions $C : \mathcal{Y} \times \mathcal{A} \mapsto \mathbb{R}$
- According to Bayesian Decision Theory, the best action to take is the one that minimizes the cost

$$a^\star = \arg\min_{a \in \mathcal{A}} \mathbb{E}_y[C(y, a)|X]$$

Overview

- Homework 4 Recap
  - Problem set 2 (Questions 4-7)
  - Problem set 5 (Questions 16-19)

- Decision Theory
  - Theory recap
  - **Examples**

Example: Medical Prediction

- Suppose that we have created a probabilistic model that determines whether a given X-ray scan contains patterns of cancer or not
- Assume that mispredictions carry a certain amount of risk, which we will fictionally model with a single scalar value
- For predicting that the X-ray scan is cancerous while the patient is healthy *(false positive), the associated cost is 10*
- On the other hand, if the prediction comes out as non-cancerous, but the patient has cancer, *the cost is 20000 (false negative)*
- Determine the action that minimizes the cost of positive prediction

Example: Medical Prediction

We have:

- label set for a non-cancerous and cancerous X-ray scan, respectively: $\mathcal{Y} = \{-1, +1\}$
- our estimated probabilistic model (e.g. a binary logistic regression):

  $P(y = -1|X) = 1 - p \quad P(y = +1|X) = p$
- Cost function, per description, highly asymmetric:

$$C(y,a) = \begin{cases} 0 & y = -1, a = -1 \\ 0 & y = +1, a = +1 \\ 10 & y = -1, a = +1 \\ 2000 & y = +1, a = -1 \end{cases}$$

Example: Medical Prediction

- From the definition of expected value, we can find the expected cost for both actions:
  - When the model predicts the X-ray to be cancerous

$$\begin{aligned}\mathbb{E}[C(y,+1)|X] &= P(y=-1|x)\cdot C(-1,+1) + P(y=+1|x)\cdot C(+1,+1) \\ &= (1-p)\cdot 10 + p\cdot 0 = 10(1-p)\end{aligned}$$

Example: Medical Prediction

- From the definition of expected value, we can find the expected cost for both actions:
  - When the model predicts the X-ray to be cancerous

$$\begin{aligned}\mathbb{E}[C(y,+1)|X] &= P(y=-1|x)\cdot C(-1,+1) + P(y=+1|x)\cdot C(+1,+1)\\ &= (1-p)\cdot 10 + p\cdot 0 = 10(1-p)\end{aligned}$$

  - When the model predicts the X-ray to be healthy

$$\begin{aligned}\mathbb{E}[C(y,-1)|X] &= P(y=-1|x)\cdot C(-1,-1) + P(y=+1|x)\cdot C(+1,-1)\\ &= (1-p)\cdot 0 + p\cdot 20000 = 20000p\end{aligned}$$

Example: Medical Prediction

- Now we can draw conclusions:
  - The scenario where we should predict the cancer is with condition:

$$\begin{aligned}\mathbb{E}[C(y,+1)|X] &< \mathbb{E}[C(y,-1)|X] \\ 10(1-p) &< 20000p \\ 1-p &< 2000p \\ 1 &< 2001p \\ p &> \frac{1}{2001}\end{aligned}$$

Example: Defining Actions

- Actions do not necessarily have to be a discrete set, but can also cover a continuous domain

- Imagine that you would like to decide on a price to sell a piece a used phone online. In this context, an action could span the whole positive real line that would determine the listing price $a \in [0, +\infty)$
  - The probabilistic model could be built based on previous phone listings and their attributes such as brand, age, condition and and finally determine the probability of a model being sold at a given price

Summary

- In order to apply the principles of decision making, we go through two stages:
  - *Learning stage*: we build the probabilistic model from labeled data
  - *Decision stage*: we use our probabilistic model, enhanced with cost function information and associated labels, to make informed decisions
- As seen in the lecture, for simpler loss functions, we can compute optimal decisions directly from the probability distributions
  - e.g. logistic or linear regression with symmetric costs
- If the costs are asymmetric, or there is a lot of uncertainty about the data, having an informed decision stage is important

References

- Lecture slides
- http://www.cs.jhu.edu/~ayuille/courses/Stat161-261-Spring14/RevisedLectureNotes2.pdf
- http://www.statsathome.com/2017/10/12/bayesian-decision-theory-made-ridiculously-simple/#examples-part-1