Spam Filtering

Salar Mohtaj | DFKI

Spam filtering

- Spam filtering task
- Naïve Bayes spam filtering
- Feed forward neural network for spam filtering
- Evaluation of spam filters

Spam detection

- Spam filtering task
- Naïve Bayes spam filtering
- Feed forward neural network for spam filtering
- Evaluation of spam filters

Spam filtering task

- A spam filter is a program that is used to detect unsolicited and ***unwanted email*** and prevent those messages from getting to a user's inbox
- The simplest and earliest can be set to watch for ***particular words*** in the subject line of messages and to exclude these from the user's inbox

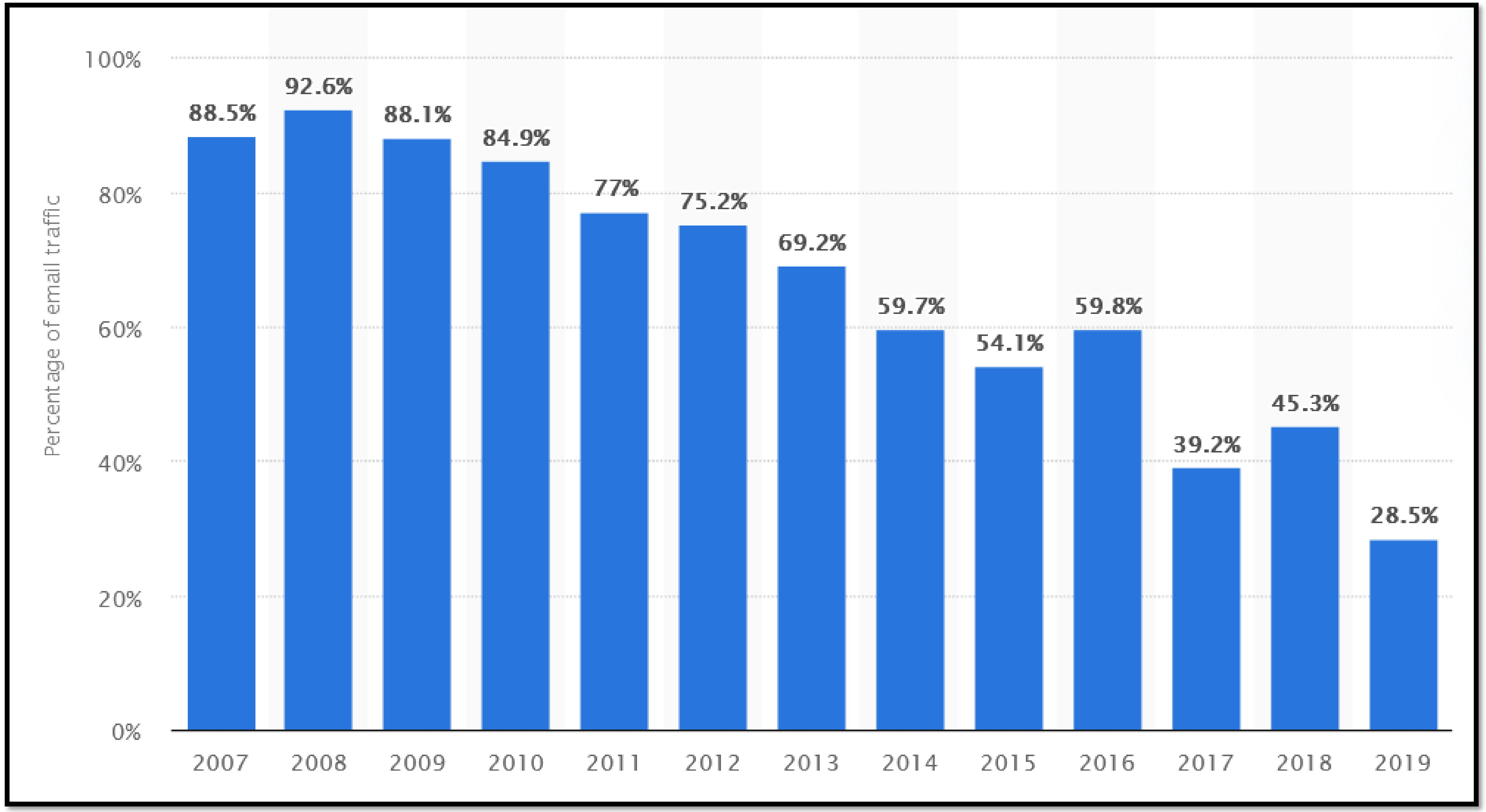
Spam filtering task

- **Content filters:** parse the content of messages, scanning for words that are commonly used in spam emails.
- **Header filters:** examine the email header source to look for suspicious information (such as spammer email addresses).
- **Blocklist filters:** stop emails that come from a blocklist of suspicious IP addresses. Some filters go further and check the IP reputation of the IP address.
- **Rules-based filters:** apply customized rules designed by the organization to exclude emails from specific senders, or emails containing specific words in their subject line or body.

type	text
ham	Hope you are having a good week. Just checking in
ham	K..give back my thanks.
ham	Am also doing in cbe only. But have to pay.
spam	complimentary 4 STAR Ibiza Holiday or £10,000 cash needs your URGENT collection. 09066364349 NOW fro...

Classification task

Spam filtering task



Graph from http://statista.com

Spam detection

- Spam filtering task
- Naïve Bayes spam filtering
- Feed forward neural network for spam filtering
- Evaluation of spam filters

Naïve Bayes spam filtering

- The Naive Bayes classifier is a simple classifier that classifies based on probabilities of events
    - It is the applied commonly to text classification
- Though it is a simple algorithm, it performs well in many text classification problems
- It is a classification technique based on **Bayes' theorem** with an assumption of independence among predictors
- As with any machine learning model, we need to have an existing set of examples (training set) for each category (spam/non-spam)

Naïve Bayes spam filtering

$$congratulations\ you\ have\ won\ a\ playstation\ 5$$

$$P(ham|congratulations\ you\ have\ won\ a\ playstation\ 5)$$

$$P(spam|congratulations\ you\ have\ won\ a\ playstation\ 5)$$

$$P(C_k|X)$$

$$C_1 = ham$$
$$C_2 = spam$$
$$X = congratulations\ you\ have\ won\ a\ playstation\ 5$$

Naïve Bayes spam filtering

$$P(C_k|X)$$

- The problem with the above formulation is that if the number of features n is large then basing such a model on probability tables is infeasible

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' theorem

$$P(C_k|X) = \frac{P(C_k)P(X|C_k)}{P(X)}$$

Naïve Bayes spam filtering

$$P(C_k|X) = \frac{P(C_k)P(X|C_k)}{P(X)}$$

naïve" conditional independence assumptions

$$P(C_k)P(X|C_k) = P(C_k)P(x_1|C_k)P(x_2|C_k) \dots P(x_n|C_k)$$

$$\hat{y} = \underset{k\in\{1,\dots,K\}}{argmax}\, P(C_k)\prod_{i=1}^{n} P(x_i|C_k)$$

Naïve Bayes spam filtering

congratulation you have won a gift card	spam
your package is out for delivery	ham
the event is postponed to the next week	ham
your PlayStation account is temporary locked	spam
congratulation you are hired	spam
congratulation you have won a PlayStation 5	**?**

congratulation you have won gift card	spam
your PlayStation account is temporary locked	spam
congratulation you are hired	spam
your package is out delivery	ham
event is postponed next week	ham

Naïve Bayes spam filtering

congratulation you have won a PlayStation 5	?

congratulation you have won a gift card	spam
your package is out for delivery	ham
the event is postponed to the next week	ham
your PlayStation account is temporary locked	spam
congratulation you are hired	spam
congratulation you have won a PlayStation 5	?

$$P(C_k)P(X|C_k) = P(C_k)P(congratulation|C_k)P(you|C_k)P(have|C_k)P(won|C_k)P(PlayStation|C_k)$$

$P(C_k)$

$P(ham) = {}^{2}/_{5}$

$P(spam) = {}^{3}/_{5}$

Naïve Bayes spam filtering

congratulation you have won a gift card	spam
your package is out for delivery	ham
the event is postponed to the next week	ham
your PlayStation account is temporary locked	spam
congratulation you are hired	spam
congratulation you have won a PlayStation 5	?

	you	account	PlayStation	is	hired	…	week	locked	congratulation
S_1	1	0	0	0	0		0	0	1
S_2	0	0	0	1	0		0	0	0
S_3	0	0	0	0	0		1	0	0
S_4	0	1	1	1	0		0	1	0
S_5	1	0	0	0	1		0	0	1

Naïve Bayes spam filtering

	you	account	PlayStation	is	hired	…	week	locked	congratulation
S_1	1	0	0	0	0		0	0	1
S_4	0	1	1	1	0		0	1	0
S_5	1	0	0	0	1		0	0	1
S_2	0	0	0	1	0		0	0	0
S_3	0	0	0	0	0		1	0	0

$$P(C_k)P(X|C_k) = P(C_k)P(congratulation|C_k)P(you|C_k)P(have|C_k)P(won|C_k)P(PlayStation|C_k)$$

$$P(congratulation|spam) = {}^{2}/_{16}$$

Naïve Bayes spam filtering

	you	account	PlayStation	is	hired	…	week	locked	congratulation
S_1	1	0	0	0	0		0	0	1
S_4	0	1	1	1	0		0	1	0
S_5	1	0	0	0	1		0	0	1
S_2	0	0	0	1	0		0	0	0
S_3	0	0	0	0	0		1	0	0

$$P(C_k)P(X|C_k) = P(C_k)P(congratulation|C_k)P(you|C_k)P(have|C_k)P(won|C_k)P(PlayStation|C_k)$$

$$P(congratulation|spam) = {}^{2}/_{16}$$

$$P(congratulation|ham) = {}^{0}/_{10}$$

Naïve Bayes spam filtering

	you	account	PlayStation	is	hired	…	week	locked	congratulation
S_1	1	0	0	0	0		0	0	1
S_4	0	1	1	1	0		0	1	0
S_5	1	0	0	0	1		0	0	1
S_2	0	0	0	1	0		0	0	0
S_3	0	0	0	0	0		1	0	0

$$P(C_k)P(X|C_k) = P(C_k)P(congratulation|C_k)P(you|C_k)P(have|C_k)P(won|C_k)P(PlayStation|C_k)$$

$$P(congratulation|spam) = {}^{2}/_{16}$$

$$P(congratulation|ham) = {}^{0}/_{10}$$

$$P(is|spam) = {}^{1}/_{16}$$

$$P(is|ham) = {}^{1}/_{10}$$

Naïve Bayes spam filtering

$$P(C_k)P(X|C_k) = P(C_k)P(congratulation|C_k)P(you|C_k)P(have|C_k)P(won|C_k)P(PlayStation|C_k)$$

$$P(ham|X) = {}^{2}/_{5} \times {}^{1}/_{10} \times \cdots$$

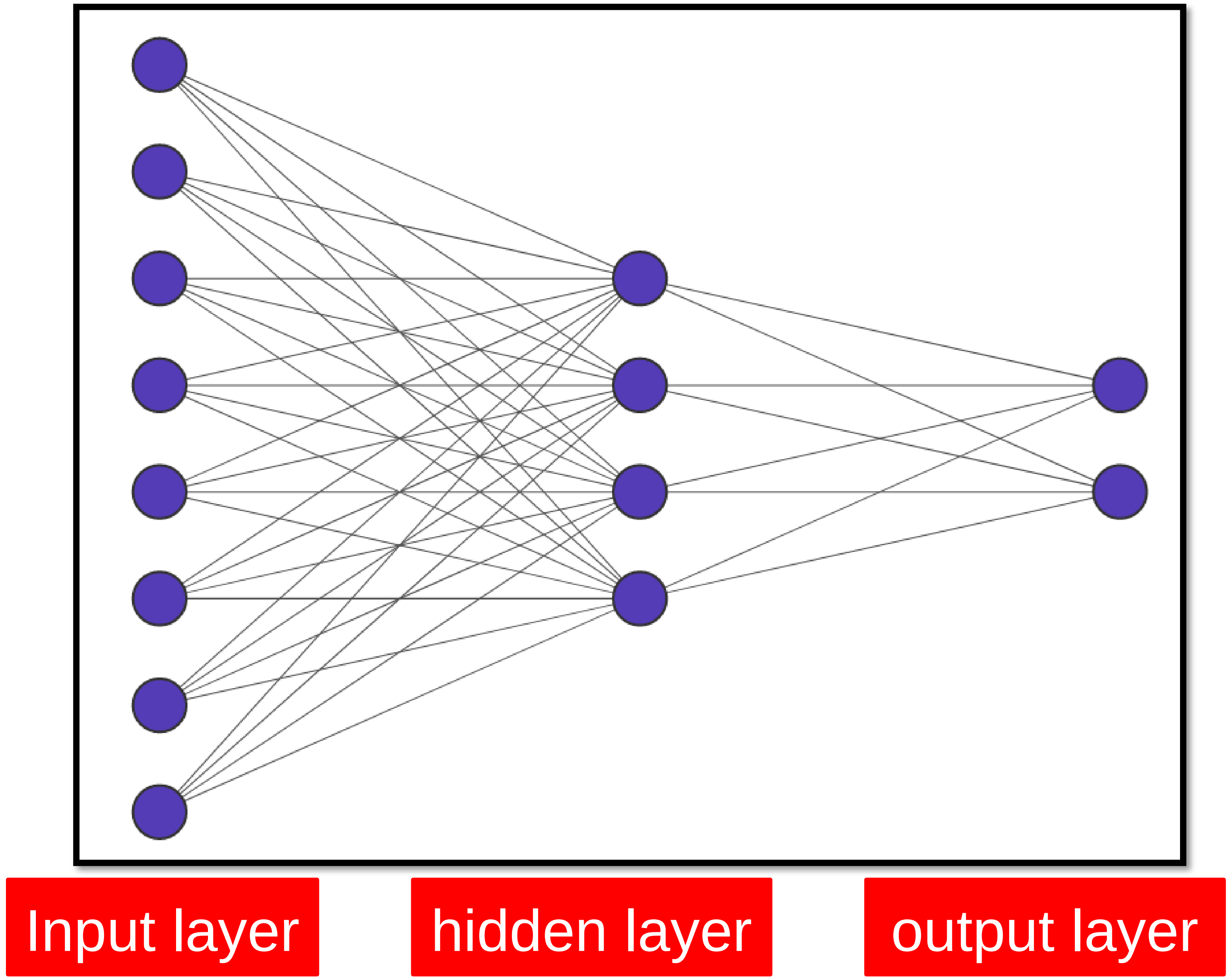
$$P(spam|X) = {}^{3}/_{5} \times {}^{2}/_{16} \times \cdots$$

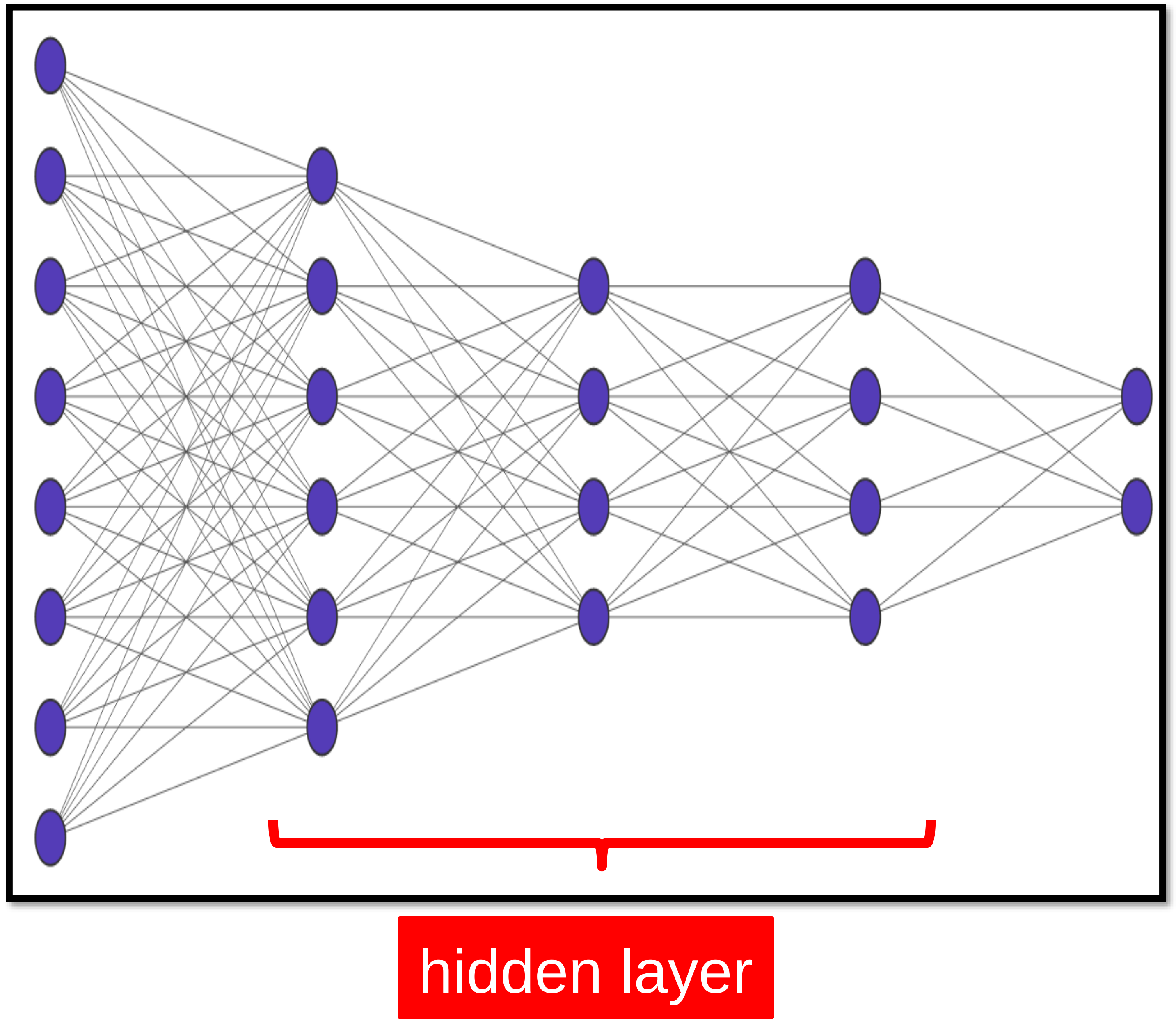
$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{argmax}\, P(C_k) \prod_{i=1}^{n} P(x_i|C_k)$$

Spam detection

- Spam filtering task
- Naïve Bayes spam filtering
- Feed forward neural network for spam filtering
- Evaluation of spam filters

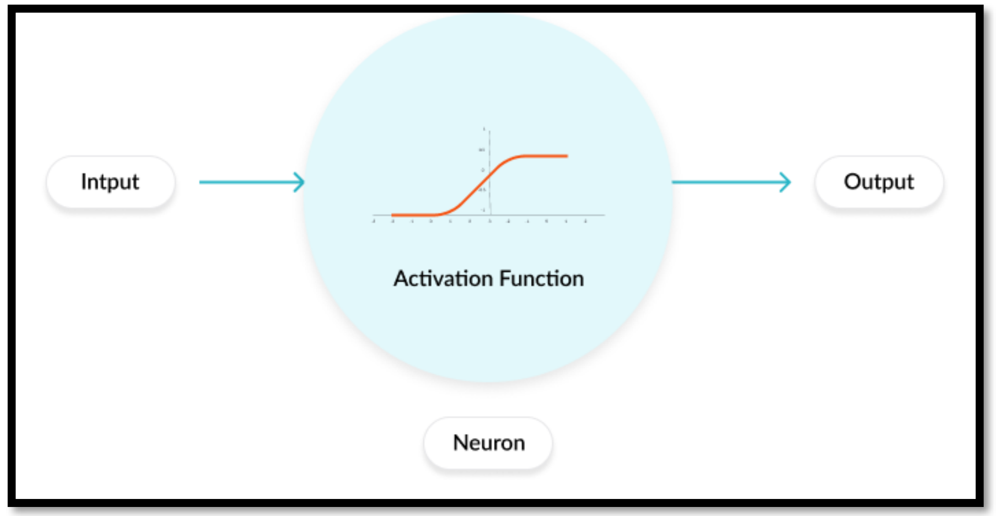
Feed forward neural network for spam filtering

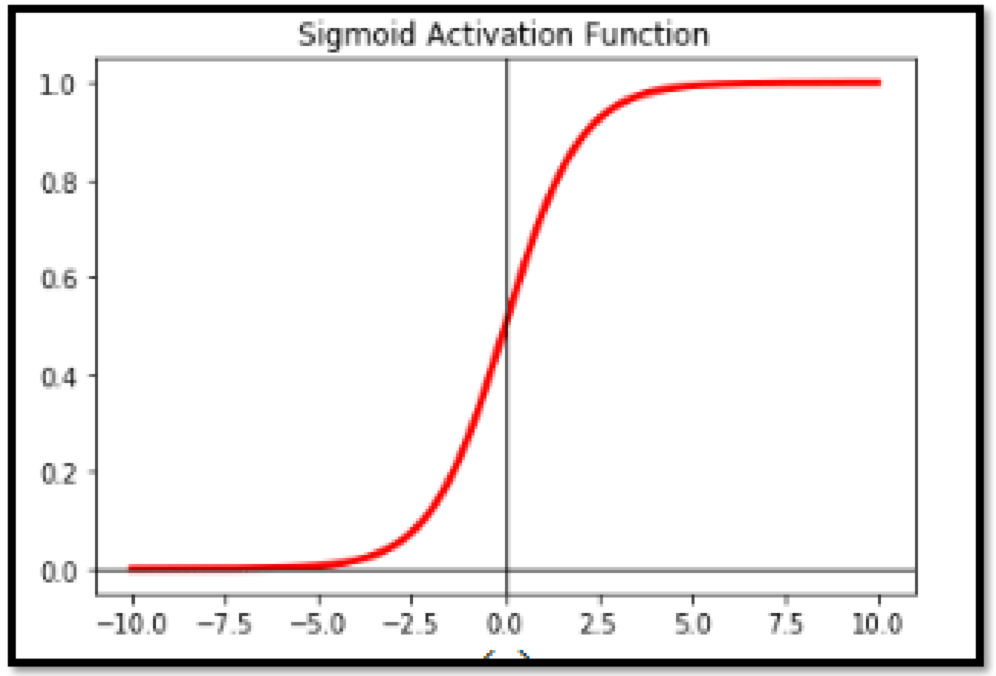


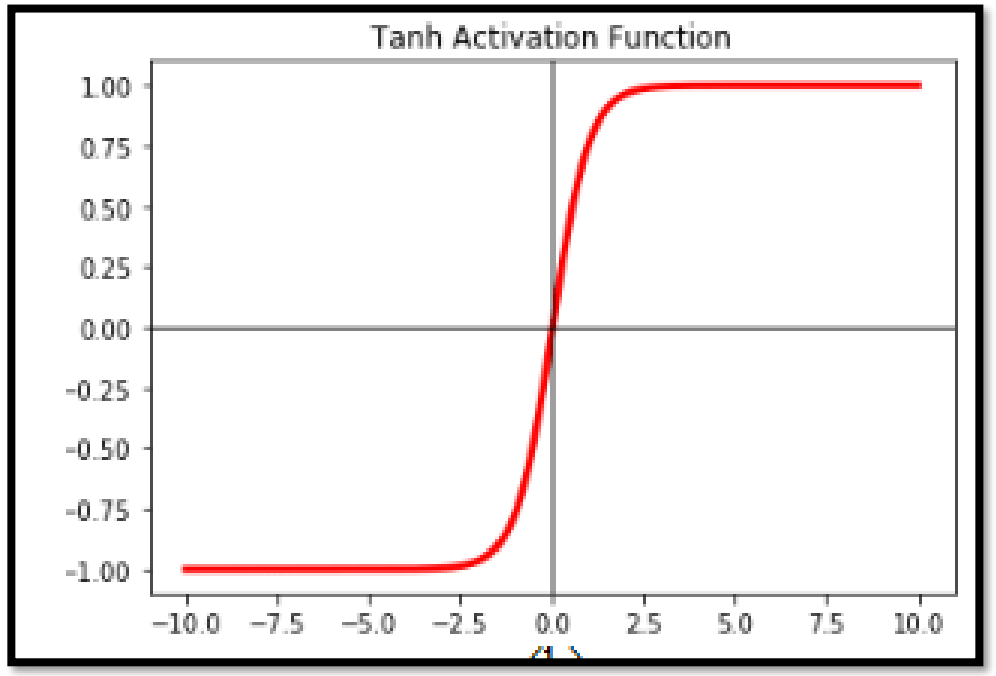


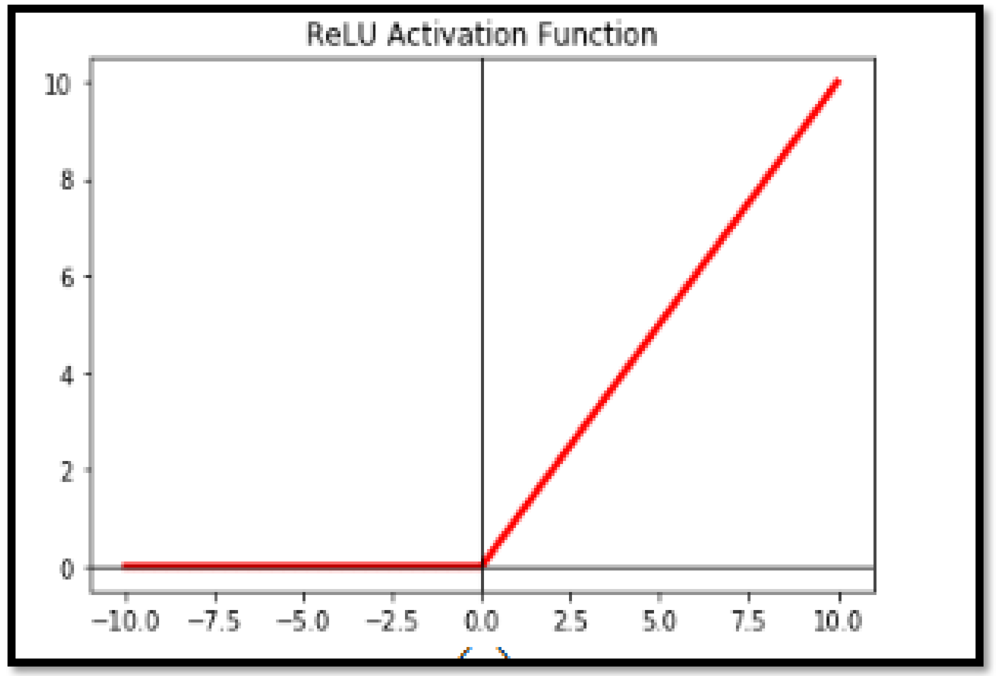
Feed forward neural network for spam filtering

- Activation function



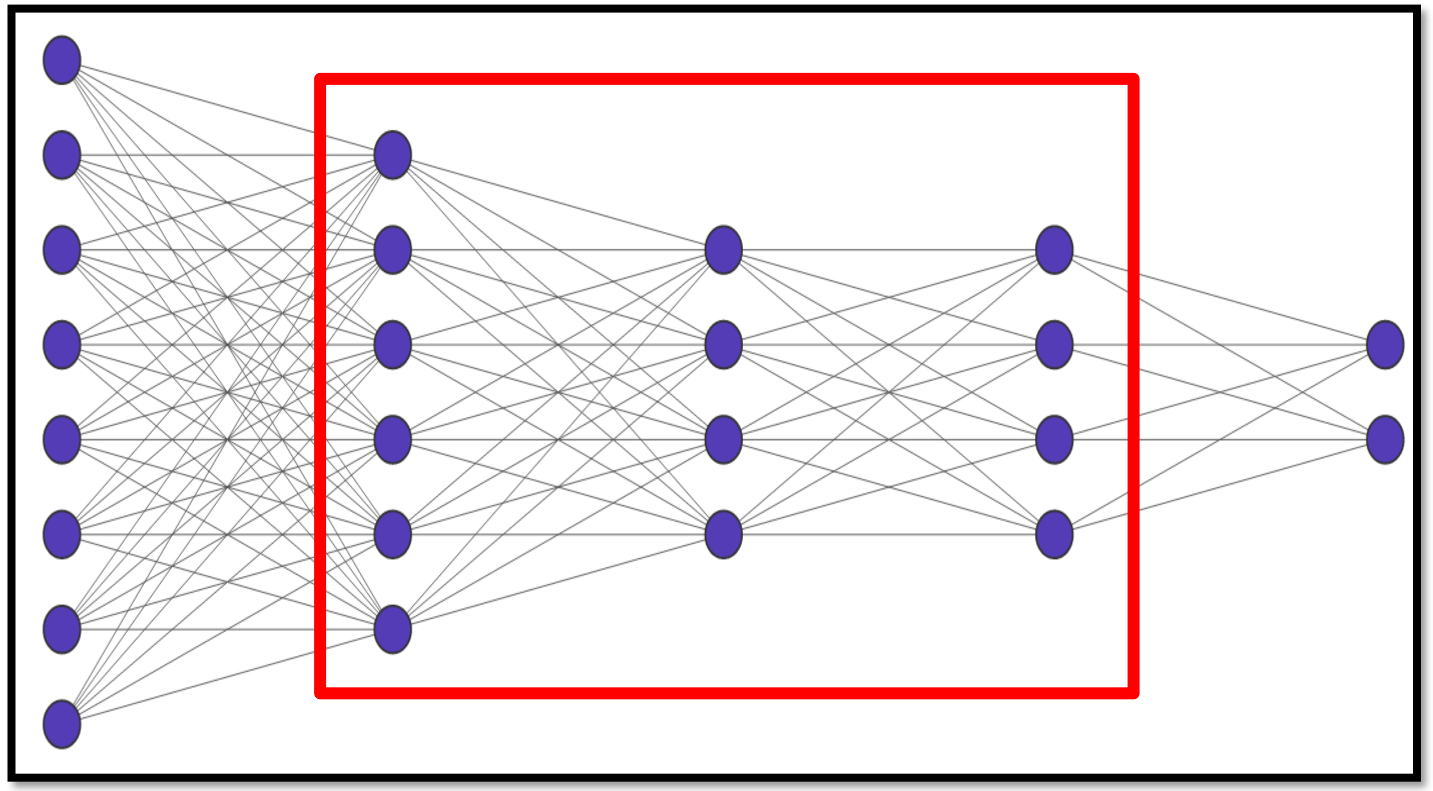


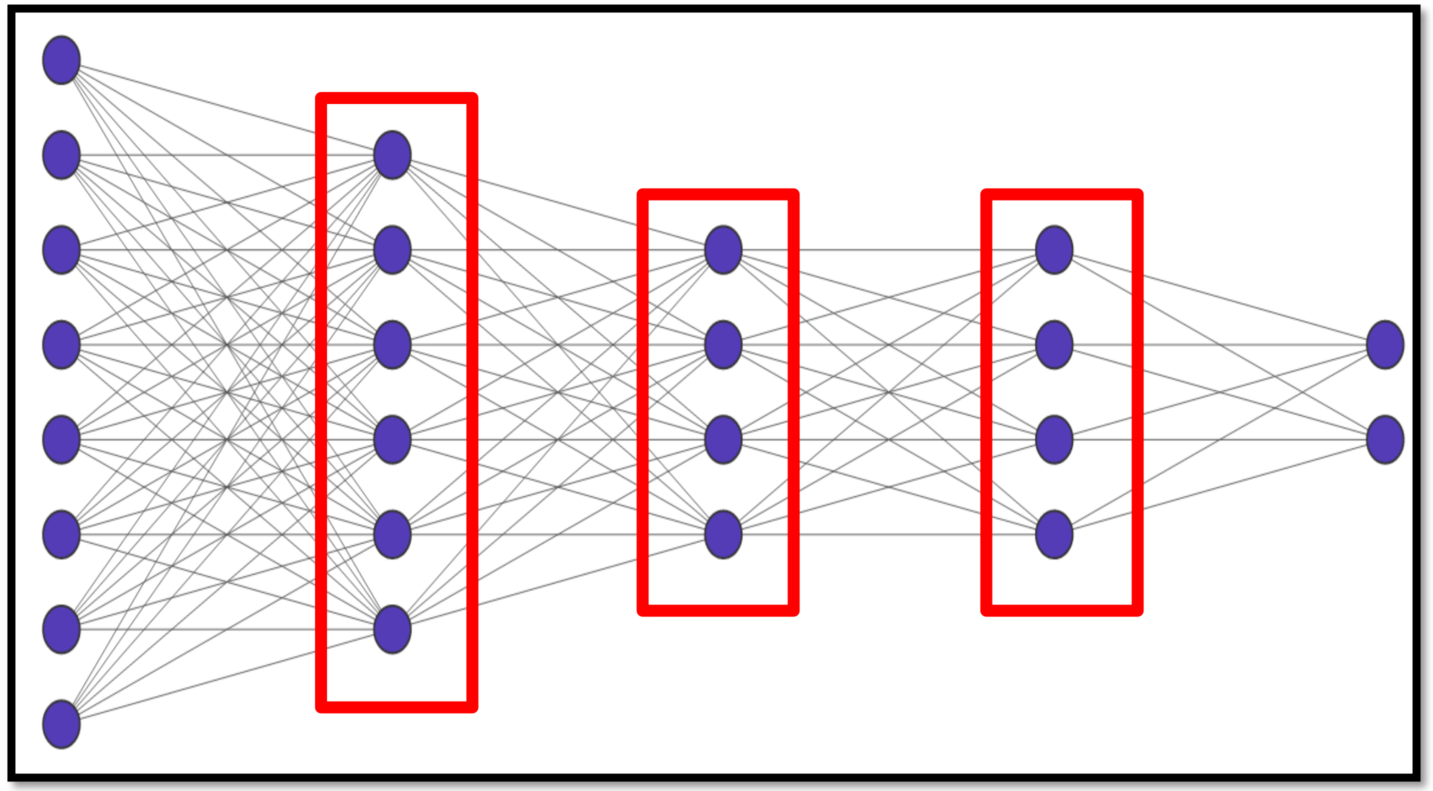




Feed forward neural network for spam filtering

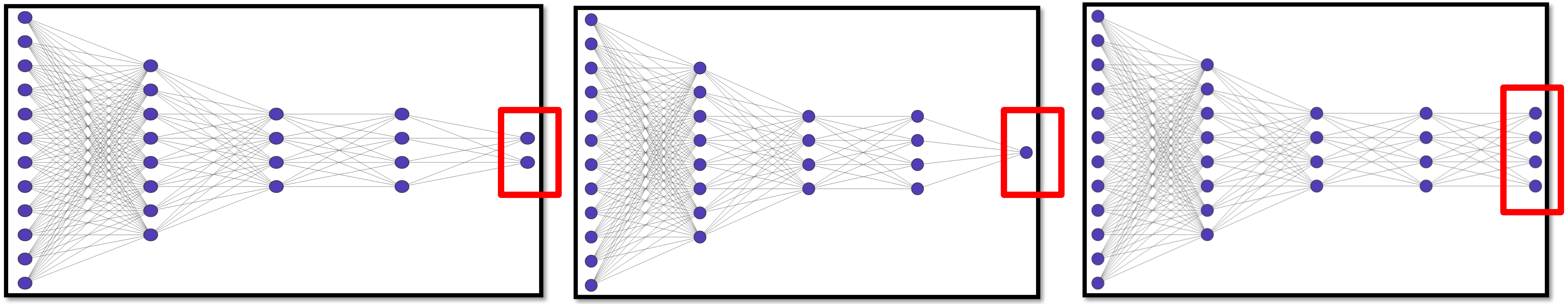
- Activation function
- Loss function
- Number of hidden layers
- Number of neurons in each layer





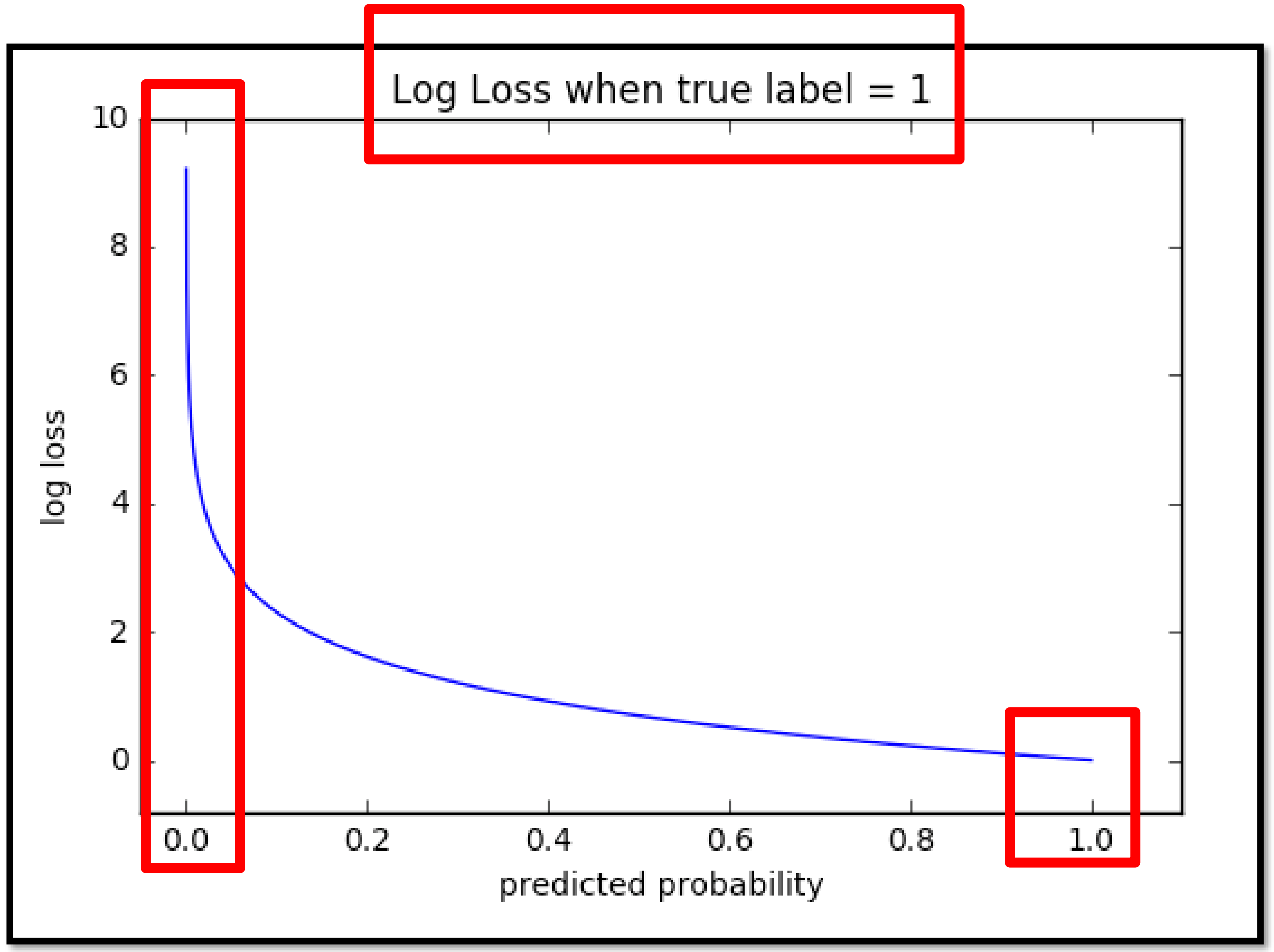
Feed forward neural network for spam filtering

- Activation function
- Loss function
- Number of hidden layers
- Number of neurons in each layer
- Output layer



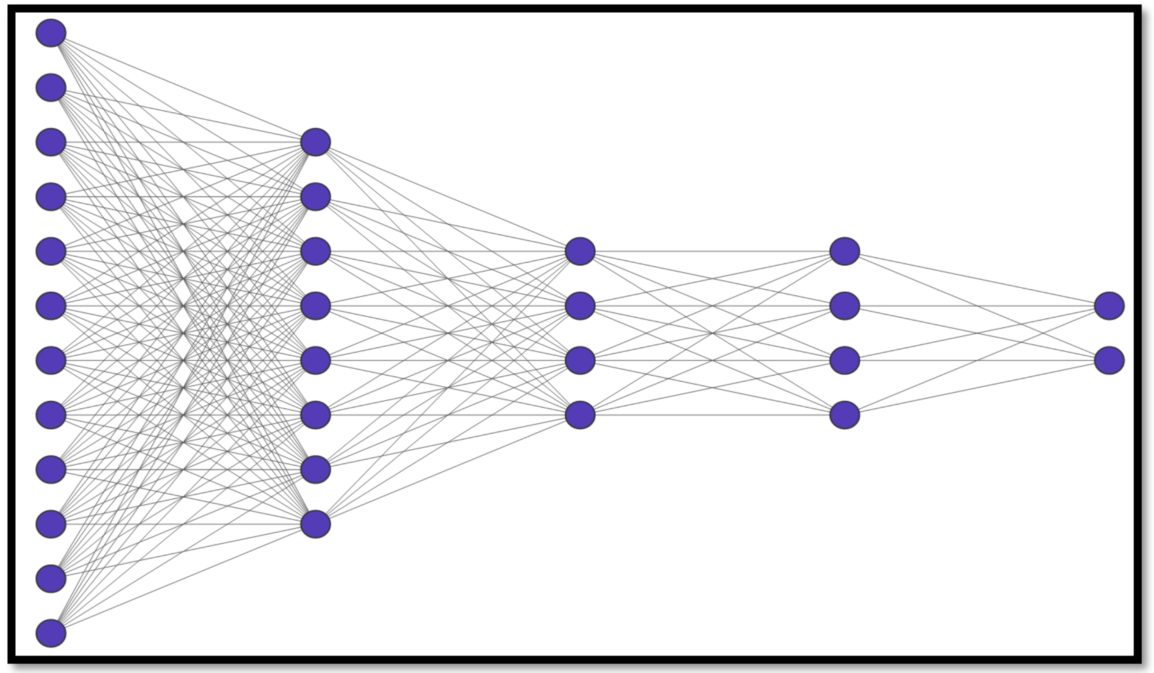
Feed forward neural network for spam filtering

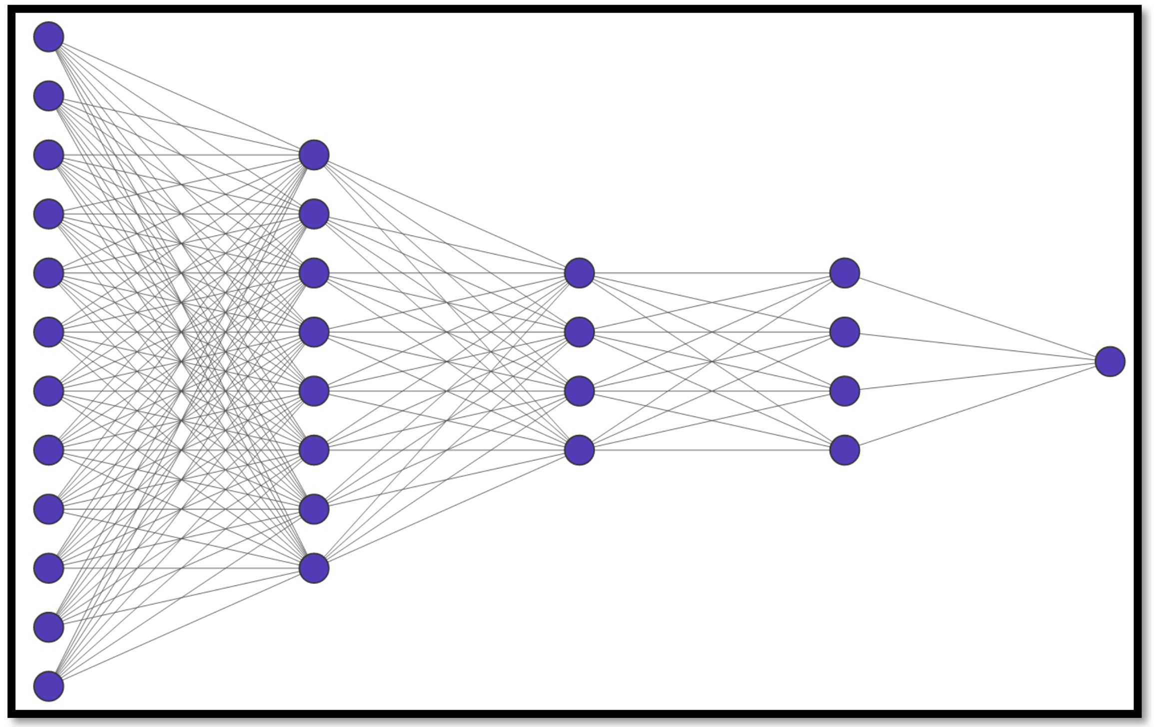
- Setting for spam filtering
  - Loss function
    - Cross entropy loss (log loss)



Feed forward neural network for spam filtering

- Setting for spam filtering
  - Loss function
    - Cross entropy loss (log loss)
  - Output layer
    - Softmax (2 output)
    - Sigmoid (1 output)





Feed forward neural network for spam filtering

- Setting for spam filtering
  - Loss function
    - Cross entropy loss (log loss)
  - Output layer
    - Softmax (2 output)
    - Sigmoid (1 output)
  - Input layer

Feed forward neural network for spam filtering

- Input layer

congratulation you have won a gift card	spam
your package is out for delivery	ham
the event is postponed to the next week	ham
your PlayStation account is temporary locked	spam
congratulation you are hired	spam
congratulation you have won a PlayStation 5	**?**

$|V|$

	you	account	PlayStation	is	hired	…	week	locked	congratulation
S_1	1	0	0	0	0		0	0	1
S_2	0	0	0	1	0		0	0	0
S_3	0	0	0	0	0		1	0	0
S_4	0	1	1	1	0		0	1	0
S_5	1	0	0	0	1		0	0	1

Feed forward neural network for spam filtering

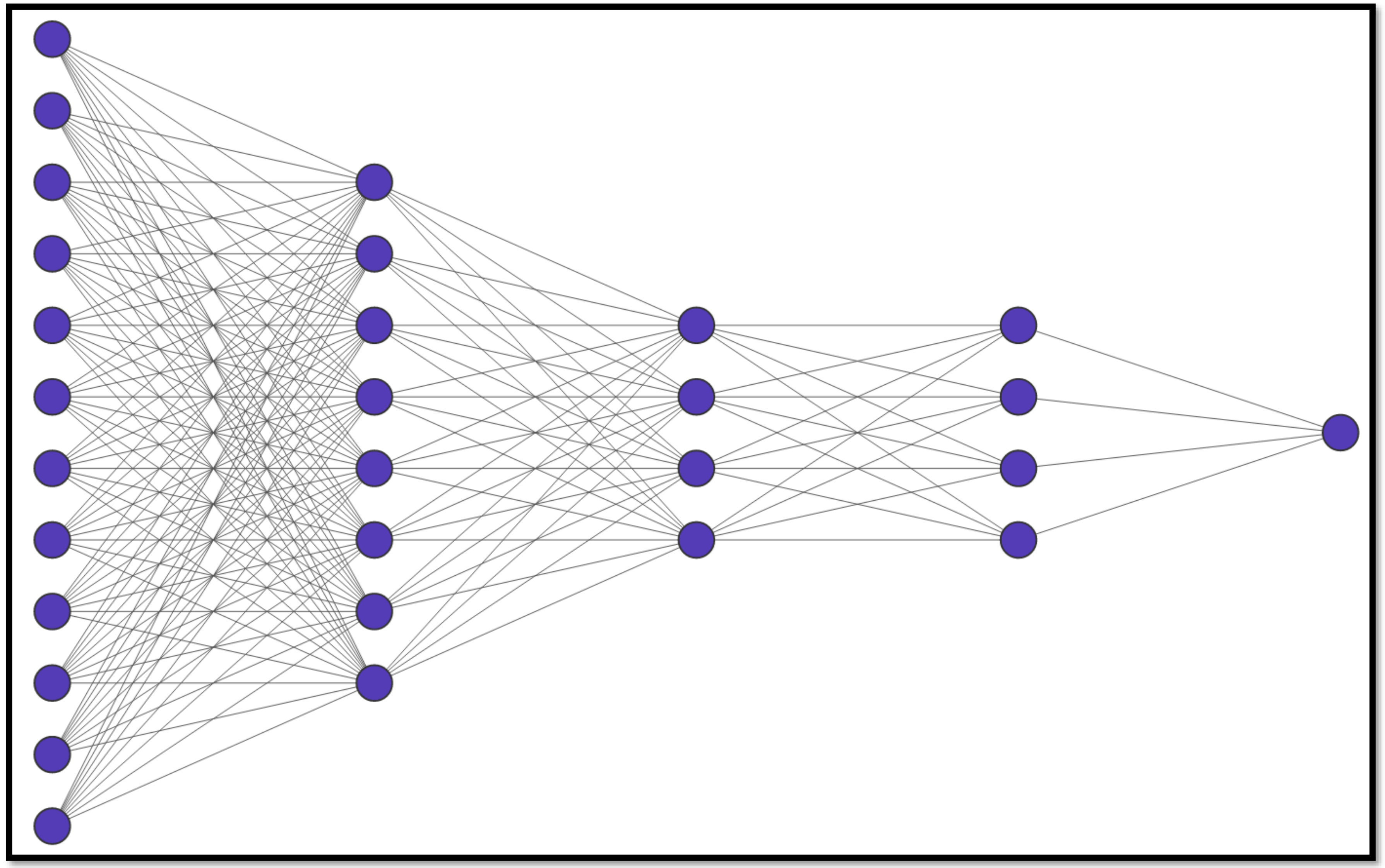
$100 - 300$

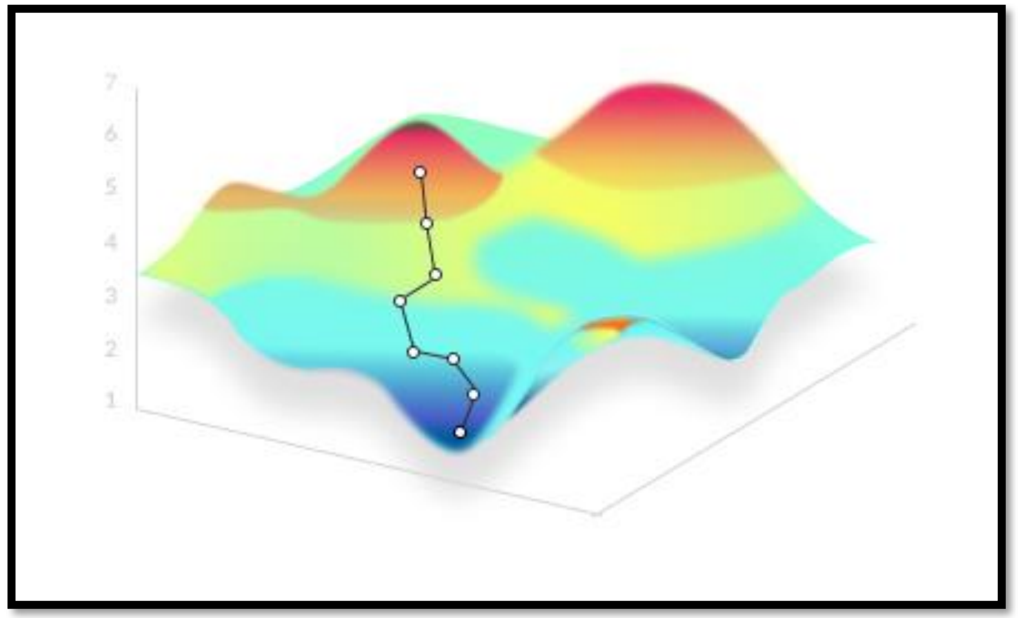
congratulation	-0.37	-0.06		0.28	-0.67
You	0.68	-0.05		0.16	0.14
Have	0.53	0.05		-0.36	-0.27
Won	0.21	-0.35		-0.53	0.20
gift	-0.81	0.41		-0.58	-0.29
card	0.14	0.01		-0.62	0.47
			Sum		
S_1	0.38	0.01		-1.65	-0.42

Feed forward neural network for spam filtering

congratulation you have won a gift card	spam

S_1
0.38
0.01
…
-1.65
-0.42





$ham/spam$

Spam detection

- Spam filtering task
- Naïve Bayes spam filtering
- Feed forward neural network for spam filtering
- Evaluation of spam filters

Evaluation of spam filters

- Confusion matrix
  - The Confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model. It is used for Classification problem where the output can be of two or more types of classes

spam → positive

ham → negative

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: True Positive	FP: False Positive
	Negative	FN: False Negative	TN: True Negative

Evaluation of spam filters

- Accuracy

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: True Positive	FP: False Positive
	Negative	FN: False Negative	TN: True Negative

Evaluation of spam filters

- Precision
- Recall
- F1

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times (precision \times Recall)}{(precision + Recall)}$$

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: True Positive	FP: False Positive
Predicted class	Negative	FN: False Negative	TN: True Negative

Evaluation of spam filters

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: 1	FP: 2
	Negative	FN: 2	TN: 3

Actual	Predicted
ham	ham
ham	spam
spam	spam
spam	ham
spam	ham
ham	ham
ham	ham
ham	spam

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: True Positive	FP: False Positive
	Negative	FN: False Negative	TN: True Negative

Evaluation of spam filters

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: 1	FP: 2
	Negative	FN: 2	TN: 3

Actual	Predicted
ham	ham
ham	spam
spam	spam
spam	ham
spam	ham
ham	ham
ham	ham
ham	spam

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: True Positive	FP: False Positive
	Negative	FN: False Negative	TN: True Negative

Evaluation of spam filters

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: 1	FP: 2
	Negative	FN: 2	TN: 3

Actual	Predicted
ham	ham
ham	spam
spam	spam
spam	ham
spam	ham
ham	ham
ham	ham
ham	spam

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: True Positive	FP: False Positive
	Negative	FN: False Negative	TN: True Negative

Evaluation of spam filters

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: 1	FP: 2
	Negative	FN: 2	TN: 3

Actual	Predicted
ham	ham
ham	spam
spam	spam
spam	ham
spam	ham
ham	ham
ham	ham
ham	spam

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: True Positive	FP: False Positive
	Negative	FN: False Negative	TN: True Negative

Evaluation of spam filters

- $accuracy = {}^{4}/_{8} = 0.5 = 50\%$
- $precision = {}^{1}/_{3} = 0.33 = 33\%$
- $recall = {}^{1}/_{3} = 0.33 = 33\%$
- $f1 = {}^{2\times 0.33\times 0.33}/_{0.33+0.33} = 0.33 = 33\%$

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: 1	FP: 2
	Negative	FN: 2	TN: 3

Evaluation of spam filters

imbalanced data

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: 0	FP: 0
	Negative	FN: 1	TN: 99

100

Actual	Predicted
ham	ham
ham	ham
...	...
ham	ham
ham	ham
spam	ham
ham	ham

99% accuracy!

Accuracy is not a good metric for imbalanced data!

Evaluation of spam filters

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: 0	FP: 0
	Negative	FN: 1	TN: 99

0% precision

0% recall

0% F1

imbalanced data

100

Actual	Predicted
ham	ham
ham	ham
…	…
ham	ham
ham	ham
spam	ham
ham	ham

Summary

- A spam filter is a program that is used to detect unsolicited and ***unwanted email*** and prevent those messages from getting to a user's inbox

$$P(C_k)P(X|C_k) = P(C_k)P(x_1|C_k)P(x_2|C_k)\dots P(x_n|C_k)$$

$$congratulations\ you\ have\ won\ a\ playstation\ 5$$

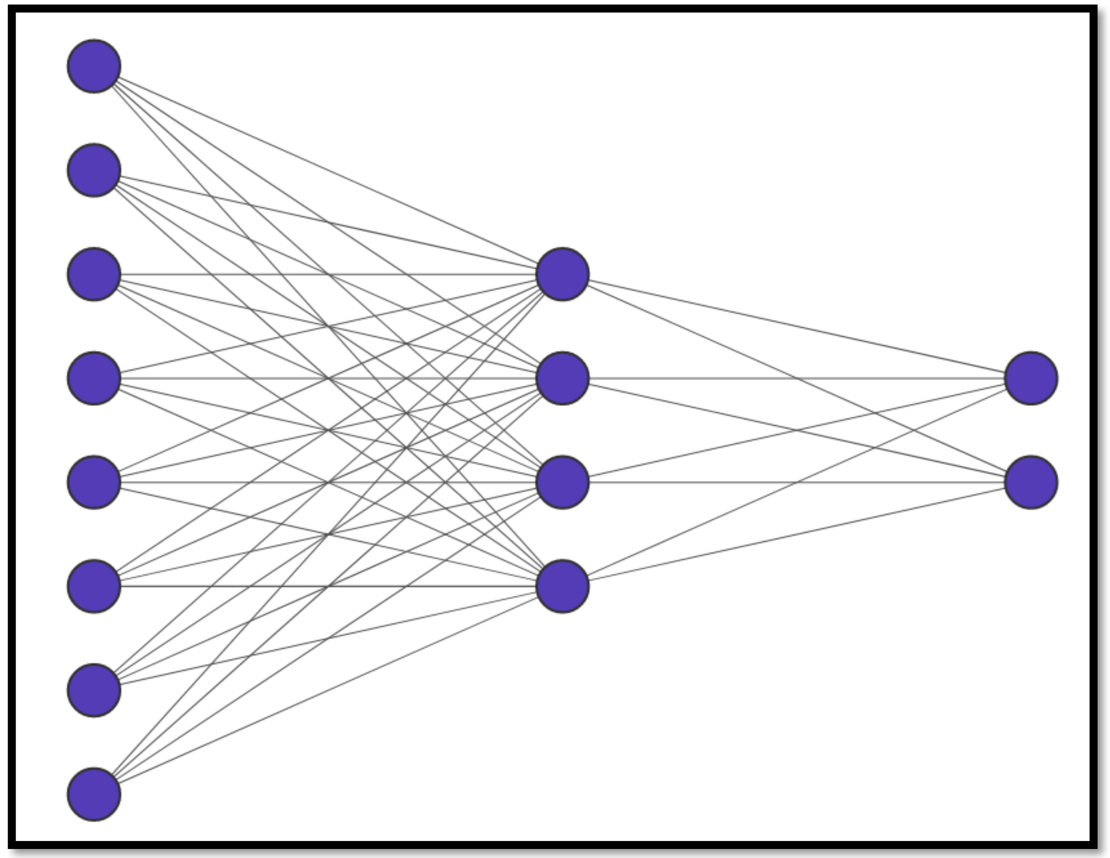
type	text
ham	Hope you are having a good week. Just checking in
ham	K..give back my thanks.
ham	Am also doing in cbe only. But have to pay.
spam	complimentary 4 STAR Ibiza Holiday or £10,000 cash needs your URGENT collection. 09066364349 NOW fro...

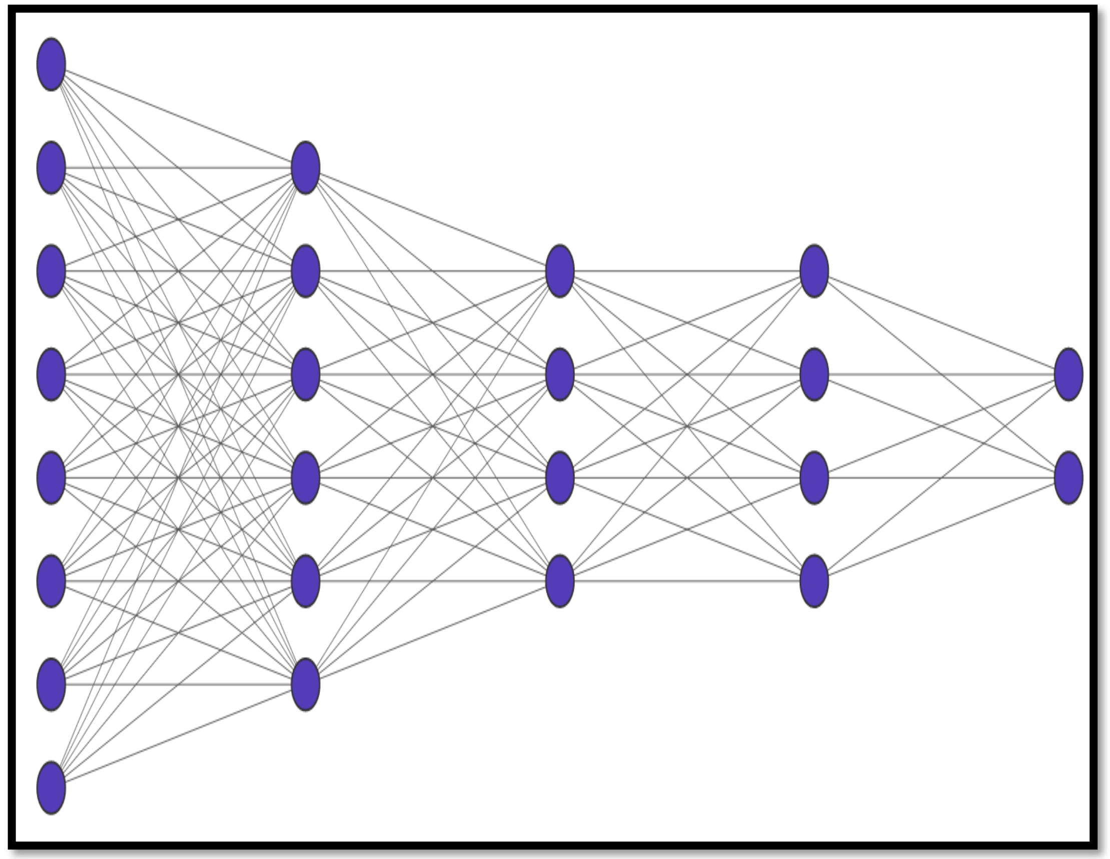
congratulation you have won gift card	spam
your PlayStation account is temporary locked	spam
congratulation you are hired	spam
your package is out delivery	ham
event is postponed next week	ham

Summary

$\underbrace{\qquad\qquad\qquad\qquad}_{}$ 100 – 300

congratulation	-0.37	-0.06	0.28	-0.67
You	0.68	-0.05	0.16	0.14
Have	0.53	0.05	-0.36	-0.27
Won	0.21	-0.35	-0.53	0.20
gift	-0.81	0.41	-0.58	-0.29
card	0.14	0.01	-0.62	0.47
		Sum		
S_1	0.38	0.01	-1.65	-0.42

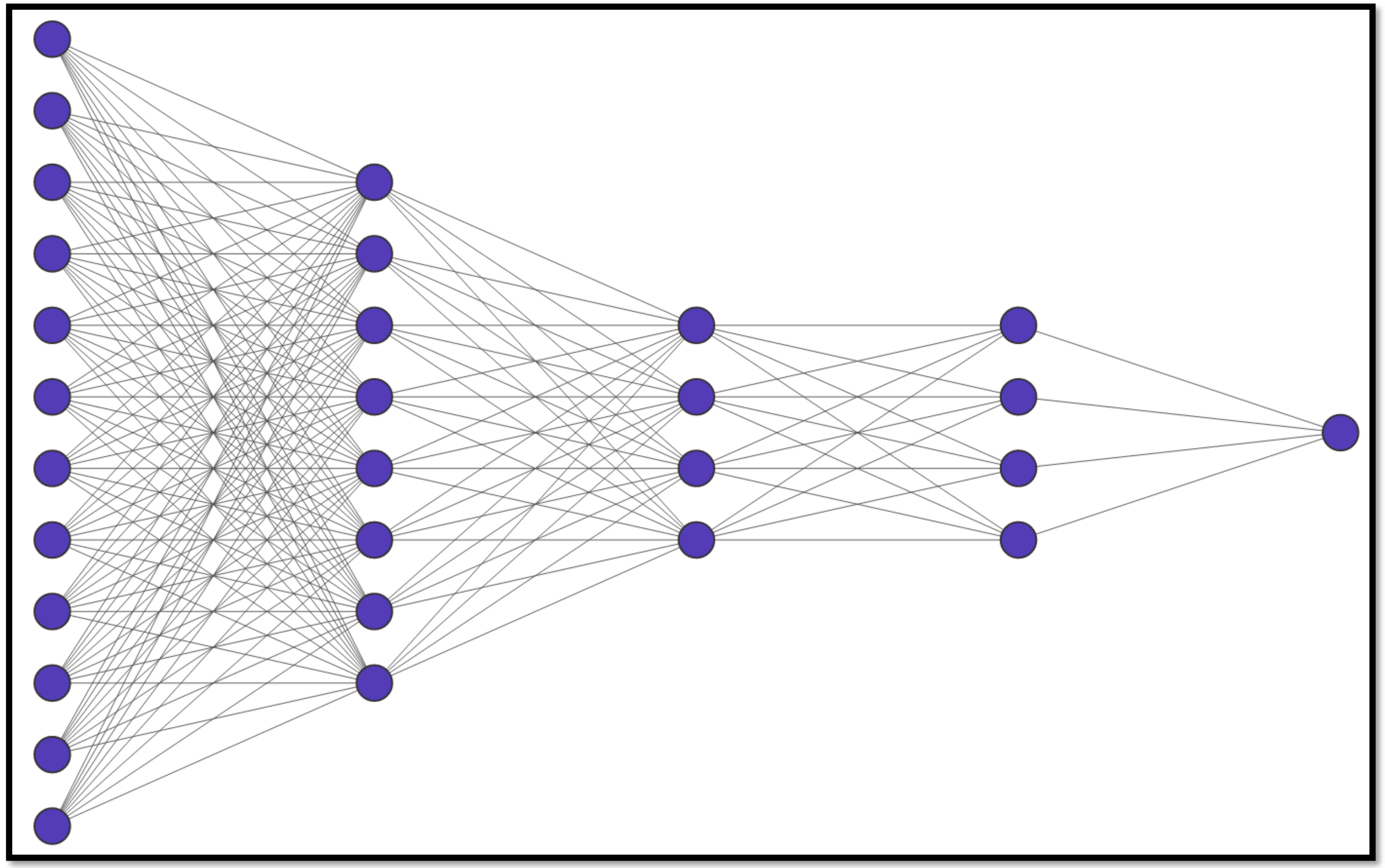


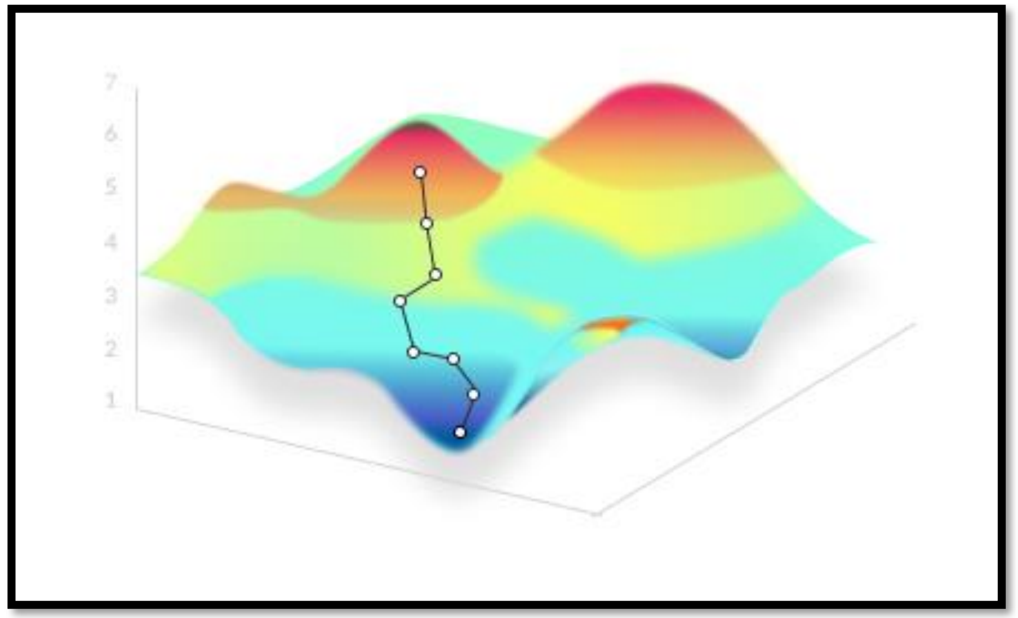


Summary

S_1
0.38
0.01
…
-1.65
-0.42

congratulation you have won a gift card spam





$ham/spam$

Summary

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times (precision \times Recall)}{(precision + Recall)}$$

		Actual class	
		Positive	Negative
Predicted class	Positive	TP: True Positive	FP: False Positive
	Negative	FN: False Negative	TN: True Negative