NLP section 5

Text Classification

Text Classification

- Text Classification is the task of assigning a label or class to a given text
- Why text classification?

  1- Spam/Not Spam
  2- Positive/Negative
  3- Fake/Real
  4- Bot/Human

Multinomial Naive Bayes: Bayesian Theorem

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

Multinomial Naive Bayes: Cont.

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d) = \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{\cancel{P(d)}} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c)P(c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c) = \underset{c_j \in C}{\operatorname{argmax}} \left[\log P(c_j) + \sum_{i \in \text{positions}} \log P(x_i \mid c_j) \right]$$

Multinomial Naive Bayes: Cont.

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Laplace for Smoothing

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} \left(count(w, c) + 1\right)}$$

$$\sum_{w \in V} \left(count(w, c) + 1\right) \qquad \left(\sum_{w \in V} count(w, c)\right) + |V|$$

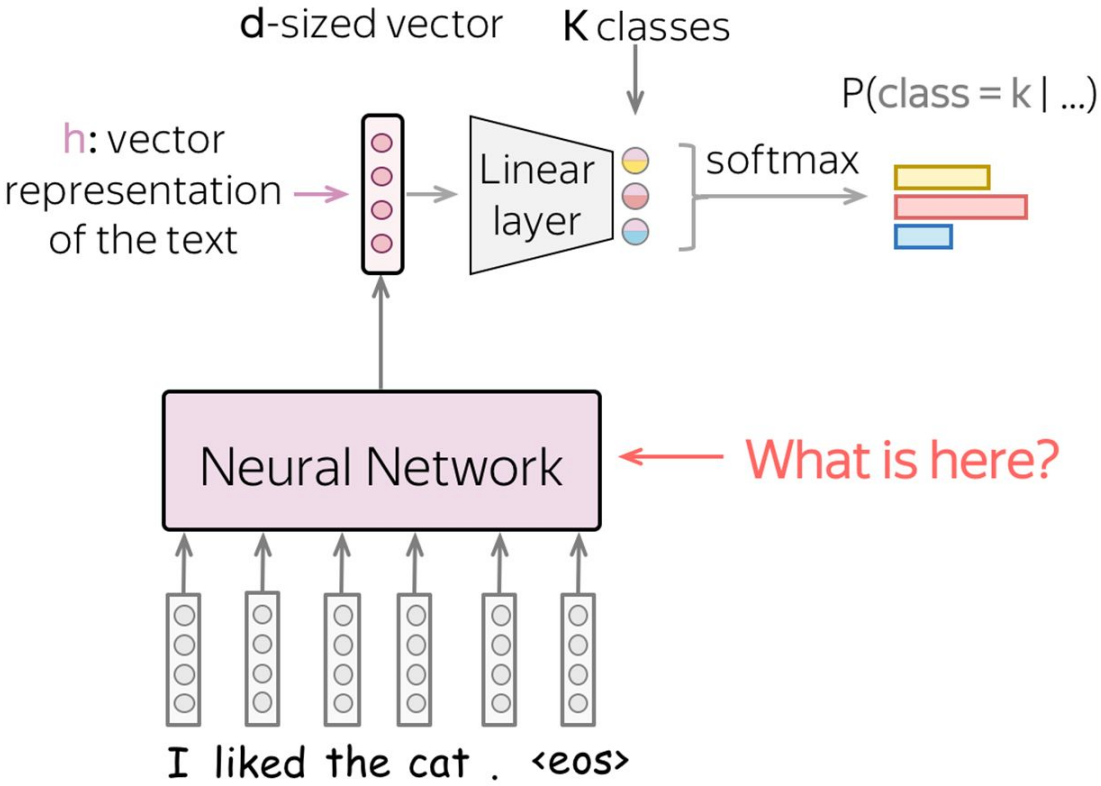
Multinomial Naive Bayes: Example

Email	Class	Content
1	Spam	"buy cheap toys now"
2	Spam	"cheap replica watches buy"
3	Spam	"buy toys cheap"
4	Ham	"meeting schedule now"
5	Ham	"project meeting tomorrow"
6	Ham	"schedule project review"

Text to Classify:

- Schedule meating review
- Buy cheap watches now

Neural Network



Code

- Multinomial Naive Bayes
- Neural Network
- Neural Network with GloVe