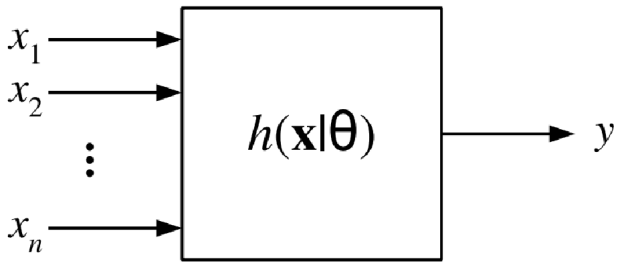
Prof. Dr. Goran Glavaš,
M.Sc. Fabian David Schmidt
M.Sc. Benedikt Ebing
Lecture Chair XII for Natural Language Processing, Universität Würzburg

11. Exercise for "Algorithmen, KI & Data Science 1"

09.02.2024

1 Introduction To Machine Learning

1. Machine learning models/algorithms aim to learn a hidden (latent) mapping between instances described with a set of variables (called features in ML) and their associated labels, in order to be able to predict the labels for instances for which the they are unknown. A general machine learning model is given in the figure below.



   a) What do the symbols in the figure denote, i.e., what are $x = (x_1, x_2, \ldots, x_n)$, y, θ, and $h(x|\theta)$?

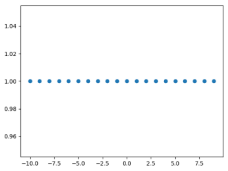
   b) Where is the "learning" in machine learning, i.e., what are we learning and how?

   c) Describe the three machine learning paradigms and explain their differences!

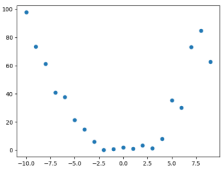
   d) What are the hyperparameters of a machine learning model? Provide an example.

   e) How do we find the optimal values for model's hyperparameters? What are underfitting and overfitting and how may we recognize them?

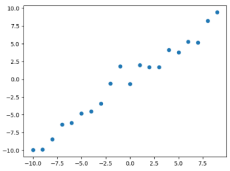
2. Imagine our model $h(x|\theta)$ is a polynomial function. For each of the given plots below, give a polynomial degree that (1) fits the data, (2) underfits the data, and (3) overfits the data. Example: A polynomial of degree 0 (i.e., constant) would fit the data in plot (a). A polynomial of degree 2 would overfit the data in plot (a)



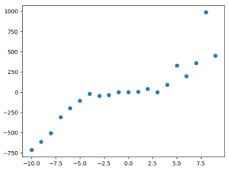
(a)



(b)



(c)



(d)

2 Naive Bayes

Clarification: treat the each word as a distinct feature like in the lecture. The only difference now is that a feature can be repeated in a single instance. Laplace smoothing here refers to "plus-1" smoothing, i.e. $\alpha = 1$.

1. Given the training examples below, compute the following probabilities (a) without and (b) with Laplace smoothing: $P(Yes)$, $P(No)$, $P(Macao|Yes)$, $P(Tokyo|Yes)$, $P(Shanghai|Yes)$, $P(Macao|No)$, $P(Tokyo|No)$, $P(Shanghai|No)$

Features	Label
Macao Tokyo Tokyo	Yes
Macao Macao	Yes
Tokyo Macao	Yes
Shanghai Tokyo Macao	No

2. Classify the following examples using the probabilities obtained (a) without and (b) with Laplace smoothing. Compute the probability for class Yes and class No for each example and predict the class with the highest probability. Round to four decimal places.

Features
Macao Macao Macao Tokyo Macao
Macao
Tokyo Tokyo Tokyo Macao
Tokyo Shanghai Tokyo Tokyo Macao
Tokyo Tokyo Tokyo

3 Logistic Regression

Logistic regression is a binary classifier, given with the following (parametrized) function:

$$h(x|w) = \sigma(-\mathbf{w^T x}) = \frac{1}{1 + \exp\left(-\mathbf{w^T x}\right))} \tag{1}$$

The binary classification is then obtained by thresholding on the value of 0.5, i.e., if $h(x|w) > 0.5$ then x belongs to class 1 (or yes), otherwise to class 0 (or no).

The *bag-of-words* approach constructs a feature vector $\mathbf{x} = [x_1, x_{...}, x_{|V|}]$ for document d from the counts of the words $x_{1,...,|V|}$ as part of the vocabulary V.

The formulation here both ignores **(a)** the bias vector and **(b)** words that are not in the the vocabulary V.

1. You are given a following toy vocabulary $\{Frodo, Daenerys, ring, dragon, fire\}$ and the following two documents: *"Daenerys's dragon spit fire"* and *"Frodo carried the ring forged in fire"*. Assuming that your current value of the parameter vector is given as

   $\mathbf{w^T} = [0.48, -1.2, 2.1, -0.32, 0.01]$

   Make the binary classification prediction for the two sentences given above (use binary bag-of-words features to represent the sentences).

2. The predictions made by a machine learning model are compared with true labels (classes) of instances in order to measure the prediction error. The prediction errors are determined by the model's loss function. The loss function of the logistic regression is the so-called (binary) cross-entropy loss, which is, for a single input instance, given as follows:

$$ce(\mathbf{x}, \mathbf{w}, y) = -\left(y \ln h(\mathbf{x}|\mathbf{w}) + (1 - y) \ln(1 - h(\mathbf{x}|\mathbf{w}))\right) \tag{2}$$

Compute the cross-entropy loss for the two sentences from (b), using the same parameter vector w as in (b), assuming that the first sentence belongs to class $y_1 = 0$, and the second sentence to class $y_2 = 1$.

$$\begin{aligned}
ce(d_1, w, y_1) &= -(0 \cdot \ln 0.18 + (1-0) \cdot \ln(1-0.18)) \\
&= -\ln(0.82) \\
&= 0.198 \\
ce(d_2, w, y_2) &= -(1 \cdot \ln(0.93) + (1-1) \cdot \ln(1-0.93)) \\
&= -\ln(0.93) \\
&= 0.073 \\
\text{ce-loss} &= 0.198 + 0.073 = 0.271
\end{aligned}$$

4 ID.3 Decision Trees

Use the ID.3 algorithm for the following questions.

1. Build a decision tree from the given tennis dataset. You should build a tree to predict `PlayTennis`, based on the other attributes (but, do not use the Day attribute in your tree.). Show all of your work, calculations, and decisions as you build the tree. How many instances does the tree classify correctly?
2. Now, build a tree using only examples D1–D7. What is the classification accuracy for the training set? What is the accuracy for the test set (examples $D8$–$D14$)? explain why you think these are the results.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Table 1: The `PlayTennis` dataset