Questions for Testing Your ML Knowledge: Module 1 and 2

2022-2023

1 General

1. What is the difference between supervised and unsupervised machine learning?
2. What is semi-supervised machine learning? Name two examples of a semi-supervised approach?
3. What is the difference between regression and classification in machine learning? Provide an example of each.
4. What is meant by feature engineering? Provide an example.
5. What is the difference between generative and discriminative machine learning:
   - (a) What is the model learning in each case?
   - (b) What consequence does this have when features correlate?
6. Name an example of a generative and of a discriminative machine learning method

2 Linear regression, logistic regression and Naive Bayes Classification

- What is the difference between linear regression and logistic regression?
  - What is each approach used for?
  - What are the main differences in their implementation?
- Consider the following formula, where x are observations and y predictions:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (f(x^i) - y^i)^2 \tag{1}$$

  - What does this formula calculate? Provide both the name and a brief explanation

- What is gradient descent? Explain its objective and procedure.

- What is meant by ‘learning rate’ and how is it used in gradient descent?

- Consider the following function:

  $$f(x) = g(\Theta^T x) = \frac{1}{1 + e^{-\theta^T X}} \tag{2}$$

  - What function is this and what is it used for?
  - Name a machine learning approach that makes use of this function

- What property does a cost function need to have in order to be guaranteed to converge?

- Consider the following function:

  $$J(\theta) = \frac{1}{m}[\sum_{i=1}^{m} y^{(i)}(-log(f(x^{(i)})) + (1 - y^{(i)})log(1 - f(x^{(i)}))] \tag{3}$$

  - What does this function provide?
  - How can we use this function to optimize logistic regression?
  - Does gradient descent converge when this function is applied?

- Consider the following function:

  $$J(\theta) = \frac{1}{m}[\sum_{i=1}^{m} y^{(i)}(-log(h_\theta(x^{(i)})) + (1 - y^{(i)})log(1 - h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\Theta_j^2 \tag{4}$$

  - What is the regularization parameter?
  - What function does it have?
  - How does it work?

- How does a naive Bayes classifier determine (i.e. calculate) which class to assign?

- What assumption does naive Bayes make that makes us call it ‘naive’?

- What is the independency assumption?

- Can feature engineering be used to capture dependencies between features in generative classifiers? Why is this a good or bad idea?

3 Support Vector Machines

1. The first SVM that was introduced was called Maximal Margin Classifier. What fundamental principle behind SVMs does this name reflect?
2. What is meant by soft margin SVM?
3. What is the role of the regularization parameter in SVM?
4. How can SVMs deal with non-linearly separable data?
5. What two settings typically need to be provided when training a SVM? What does each of them do?
6. The most commonly used kernels are the linear kernel and the Gaussian kernel. What are the typical scenarios for each of these?

4 Hidden Markov Models

1. What is a first-order markov model?
2. What algorithm is used to identify the most likely state at a given position in a sequence? Provide a brief explanation of how it works.
3. What is the Viterbi algorithm?
   (a) What is it used for?
   (b) How does it work?
4. How can a HMM markov model be learned from untrained data?

5 Conditional Random Fields

1. What (potential) shortcomings of HMMs can be addressed by conditional random fields?
2. True or False? Conditional random fields are specifically designed for modeling sequences. Provide a brief explanation of your answer.
3. What algorithm can be used for optimizing conditional random fields?

6 Feature Representation

1. What is meant by feature engineering? Provide examples of steps involved.
2. Two features in a machine learning algorithm interact, what can you do when:

- These features correlate (e.g. a name being present on a gazeteer and being capitalized)?
- The presence of one feature changes the meaning of the other (e.g. the grammatical function in a passive sentence)?

Make sure to include both possibilities through feature engineering and choice of machine learning approach in your answer.

3. What is meant by a one-hot vector representation?

4. What information can be captured by dense vector representations? When does such a representation provide an advantage over one-hot vectors?

5. Suppose you have used SVMs with one-hot vector representations of words with moderate success for a classification task. You now have access to 100 dimensional, high-density word embeddings.

   (a) How can this change in input representation impact your results?

   (b) Do you think the same settings in your machine learning set-up would yield the best results? Why (not)?

6. Can part-of-speech tags be represented by lower dimensity dense vectors (instead of one-hot vectors)? What impact on the outcome of your machine learning may this have?

7. How can chunks be represented as input for machine learning? Explain the challenges involved.

8. How can syntactic dependencies be represented as input for machine learning? Explain the challenges involved.

7 Word Embeddings

1. What is pointwise mutual information? Explain what this captures when used to create word embeddings.

2. How do word embeddings created by the positive pointwise mutual information scores of their context words compare to one-hot embeddings in terms of (a) dimensity, (b) sparcity and (c) capacity of generalization.

3. What method is used to reduce the dimension of PPMI word embeddings?

4. How do word embeddings created applying PPMI and SVD compare to one-hot embeddings in terms of (a) dimensity, (b) sparcity and (c) capacity of generalization.

5. How do word embeddings created using a learning method such as word2vec compare to one-hot vectors of (a) dimensions, (b) sparcity and (c) capacity of generalization.

6. Name and briefly explain two distinct methods for creating high-density word embeddings. They may involve more than one step.
7. What is CBOW? Explain what it stands for and provide the formula.
8. How can word embeddings be evaluated?