Jurafsky: Ch-5,6,7,9, + MIT RNN Slides

Word2vec, Glove, Gensim - StanfordUniLec-1

Ch5_Jurafsky_Logistic Regression

Generative and Discriminative Classifiers: explain with example.

What are the Components of a probabilistic machine learning classifier?

Sigmoid(logistic) function: Explanation, Advantages / disadvantages-how to overcome.

Define a decision boundary for classifiers. Explain the decision boundary of a sigmoid function for a two class classifier.

Sentiment classification Example problem.

Designing features

What is representation learning?
In order to avoid the extensive human effort of feature design, recent research in NLP has focused on representation learning: ways to learn features automatically in an unsupervised way from the input. Also Ref to Chapter 6 and Chapter 7.
vector semantics model learn representations of the meaning of words directly from their distributions in texts; is an example of representation learning.

Mention the advantages Logistic regression has over naïve Bayes.
Explain how you would choose a classifier between a Logistic regression and a naïve Bayes.

The cross-entropy loss function
Explanation, derivation of formula, compute LCE for the Sentiment classification problem.

Stochastic gradient descent
Explain: "For logistic regression, this loss function is conveniently convex. By contrast, the loss for multi-layer neural networks is non-convex, and gradient descent may get stuck in local minima for neural network training and never find the global optimum."
Define learning rate in the light of gradient descent algorithm. Explain how the learning rate influences the process of finding the global optimum.
[end of 5.4.2]The learning rate h is a parameter that must be adjusted. If it's too high, the learner will take steps that are too large, overshooting the minimum of the loss function. If it's too low, the learner will take steps that are too small, and take too long to get to the minimum. It is common to begin the learning rate at a higher value, and then slowly decrease it, so that it is a function of the iteration k of training; you will sometimes see the notation hk to mean the value of the learning rate at iteration k.

Derive the Gradient Equation for Logistic regression.(5.4.1+5.8).

Write The Stochastic Gradient Descent Algorithm [Fig-5.5]

Show the computation (of a single step) of the gradient descent algorithm for an example [5.4.3]

Explain stochastic, batch training, and mini-batch training to compute the gradient over a dataset. [5.4.4]

Overfitting, generalization, and regularization

Define the softmax function for multinomial logistic regression. [5.6]

Explain why Multinomial logistic regression uses the softmax rather than the sigmoid classifier?
[the sigmoid classifier is used in binary logistic regression.]

More questions will be added from 5.6 and 5.7 in future.[may not this time]

Ch6_Jurafsky_ Vector Semantics

State the "distributional hypothesis" and explain how it leads to "vector semantics" model.
Lemmas and Senses of words.
Example of homonymous (have multiple senses) words.
Relationships between words or senses
Synonyms, antonyms
propositional meaning
the "principle of contrast" in semantics
Word Synonyms vs Word Similarity vs Word Relatedness(association)
Semantic Frames and Roles
Taxonomic Relations: hyponym (Subordinate), hypernym (Superordinate)
lexical fields and frames,
Connotations(affective meaning) and sentiment; how are they related?
Connotation → evaluation → sentiment
three important dimensions of affective meaning : valence, arousal, and dominance; explain with examples
Famous philosopher Ludwig Wittgenstein suggested that "the meaning of a word is its use in the language". Explain.
Describe the vector semantic model with some examples.
Explain how the Vector semantics model combines two intuitions: the distributionalist intuition
(defining a word by counting what other words occur in its environment), and the

vector intuition (representing each word as a point in a multi-dimensional space).
What is embedding in the vector semantics model?
two most commonly used models in the vector semantics are tf-idf model, often used as a baseline, and, the word2vec model. Compare and contrast (similarity and differences). (sparse vector vs dense vector).

What is a co-occurrence matrix? How does it help in building a vector semantic model?
What is a vector space model? How is it related with a term-document matrix?
Using suitable examples explain how documents can be represented as vectors in a vector space.
Understand fig-6.2, 6.3

Fig-6.4 visualization of the document vectors [6.3.1]

What is a term-context matrix (term-term matrix, word-word matrix)?[6.3.2]

Problem: compute cosine similarity using word-vectors:
a) raw count b) tf-idf

Problem: given a chunk of text make a word-vector. (e.g Fig-6.5; remember that Fig-6.5 does not exactly match the chunk of text given. instead Fig-6.5 is taken from Brown corpus)

Pointwise Mutual Information (PMI) Fig-6.9-6.12
PPMI ; Laplace smoothing Fig-6.9-6.12

Word2vec

What are dense vectors and sparse vectors?
Dense vectors work better in every NLP task than sparse vectors. Explain why.

Skip gram with negative sampling, sometimes called SGNS.
SGNS vs GloVe
word2vec is a much simpler model than the neural network language model, in two ways. Mention those two ways with short explanations.

Mention the four intuitive steps in the skip-gram algorithm for word-embedding.

Ch7_Jurafsky_Neural Nets and Neural Language Models

Compare and contrast Neural Networks and Logistic Regression.
Explain an artificial neural unit with necessary diagrams and formulas.

What is an activation function in a neural network.
Explain each of the three activation functions: the sigmoid, the tanh, and the rectified linear ReLU
Pros and cons of each of the three activation functions.
Draw the truth tables of simple logical functions of AND, OR, and XOR with two inputs.
Demonstrate that some of the simple logical functions of AND, OR, and XOR need multi-layer networks. [Fig-7.5]
Show that XOR is not a linearly separable function. [Fig-7.5]
Draw a neat diagram of a multilayer NN to demonstrate the functionality of a XOR logic. [Fig-7.6]
[Fig-7.7] Draw a two dimensional space representation of the inputs of a XOR gate and explain how this is transformed by a hidden layer.
What is a Feed-Forward Neural Networks. Explain using diagrams. [Fig-7.8]

The output of a hidden layer in a FF NN is given by: $h = \sigma(Wx+b)$; Explain each of the parameters using a diagram.
[Hint: draw FFNN diagram [Fig-7.8] then explain;
h is a hidden layer given by a vector $h_1, h_2, \ldots, h_j, \ldots h_{n1}$; i.e. n1 number of cells(hidden units) in this particular layer (this is the hidden layer no-1 (first layer) hence n1).
sigma (σ) is the activation function of the hidden layer h,
W is a weight matrix [wij]; i = no. associated with input xi; j = no. associated with hidden unit h_j;
x is input vector as $x_1, x_2, \ldots, x_i, \ldots x_{n0}$; [input is assumed as layer no.-0 hence n0]

Define the softmax function. Explain its significance in a neural network classifier.
Write the necessary formulas to represent a 3-layer FF net. [eqn-7.12]. should explain each of the notations.
"A feedforward neural net is an instance of supervised machine learning", explain. [beginning of 7.4]
What do you mean by training a neural net?
Define the cross-entropy loss. Why it is important in a NN?
Explain the gradient descent optimization algorithm for a FF NN [or for the logistic regression-ch-5].
What is a one-hot vector. How does it help a NN classifier?
Show that the cross-entropy loss / negative log likelihood loss in NN is given by:

$$L_{CE}(\hat{y}, y) = -\log \frac{e^{Z_i}}{\sum_{j=1}^{K} e^{Z_j}} \quad \ldots \text{ [eqn-7.16]}$$

What is a computation graph? Assume the inputs a = 3, b = 1, c = 1. Consider computing the function $L(a, b, c) = c(a + 2b)$. Showing the forward and backward pass compute the derivatives: $\frac{\partial L}{\partial a}, \frac{\partial L}{\partial b}, \frac{\partial L}{\partial c}$. [Fig. 7.9 - 7.10].

Using an example explain how a computation graph helps in finding the Backward differentiations. [Fig. 7.9 - 7.10].

7.5 Neural Language Models

Mention the advantages-disadvantages of neural net-based language models over the n-gram language models. [7.5]

What is a feedforward neural LM?

Using a suitable diagram explain a FFNNLM (FeedForward Neural Network Language Model) with window size of 3. Assume that you have an embedding dictionary E that gives the embedding for each word in a vocabulary V. [Fig. 7.12; Fig-7.13 in 2020 version]

Ch9_Jurafsky_RNN + MIT RNN slide

What is the main difference between a Feed-Forward Neural Networks (FFNN) and a Recurrent Neural Network (RNN). Explain using diagrams. [Fig-7.8 and Fig-9.2-9.3 Find a simplified fig of RNN] [MIT RNN slide-26-to-32]

Explain standard RNN gradient flow. [MIT Slide-57]
What do you mean by exploding and vanishing gradient? [MIT Slide-58-59]
Why are vanishing gradients a problem? Mention some techniques to over the problem.

Using a diagram explain how a LSTM works. [73 onwards]
Mention some popular application of RNN. [83 onwards]