Neural Networks Fall 2020

Gal Chechik, Shauli Ravfogel

Exercise 5

1. **(20pts)** *Multivariate Normal Distributions*
   a. Read this link discussing visualization of the Multivariate Normal Distributions. Using the code presented there, plot a bivariate normal distribution with various values of the covariance matrix Submit 3 different examples.
   b. Prove that for a bivariate normal distribution $N(\mu, \Sigma)$ that obeys $\Sigma_{11} = \Sigma_{22}$, and $\Sigma_{12} > 0$, the major axis of the ellipse describing the distribution follows the $45°$ line. Also prove that it is perpendicular to that line if $\Sigma_{12} < 0$.

2. **(30pts)** *Principal Component Analysis*
   a. (20 pts) In this exercise, we prove that the linear projection onto an d-dimensional subspace that maximizes the variance of the projected data is defined by the top d eigenvectors of the data covariance matrix C (the vectors corresponding to the d largest eigenvalues). The result for the case of d = 1 was proved in class.

   Prove by induction: assume the result holds for some general value d, and show that it consequently holds for dimensionality $d+1$. To do this, first compute the derivative of the total variance E(y^2 of the projected data $y \in R^{d+1}$ with respect to a vector w_{d+1} and set it to zero. This vector defines the new direction in data space. The derivative should be computed subject to the constraint that w_{d+1} be orthogonal to the previous eigenvectors vectors $w_{1,\ldots,} w_d$, and also that is be normalized to unit length. Show that the new vector w_{d+1} is an eigenvector of C. Finally, show that the variance is maximized if the eigenvector is chosen to be the one corresponding to eigenvalue λ_{d+1} where the eigenvalues have been ordered in decreasing value.

   b. (10 pts) Given an input data $x \in R^n$ with zero mean, and a linear projection matrix $W(y = Wx)$ which projects data onto a d-dimensional subspace, defined by the d largest eigenvectors of the covariance matrix of x, show that the total variance $E(y^2) = \sum_{i=1}^{d} \langle y_i^2 \rangle$ is invariant under an orthonormal rotation of the axes within that subspace: $W \rightarrow WO$, where O is an orthonormal matrix. Hint: Use the commutative property of the Trace of a matrix *trace(AB) = trace(BA)* for any two matrices *A, B*

3. **(50 pts)** *Word embeddings*

In this exercise you will implement the GloVe model and test it on a text corpus. Read the GloVe paper[1], answer the following questions and then follow the implementation instructions below:

a. In what sense, GloVe is global compared to CBOW? What is the disadvantage of CBOW compared GloVe in this respect?

b. Eqn 8 in the paper describes adding a weighting to the cost function. Explain the problem that the weighting is aimed to address..

c. Stochastic Gradient Descent (SGD). Approximate the gradient of the cost function using a single element of the co-occurrence matrix. You should write explicitly the resulting gradients and the update rules for all model parameters.

Implementation instructions:
A notebook with a skeleton for your code is available here.

i. We will use co-occurrence counts collected from 250,000 sentences in Wikipedia. The function *load_cooccurrences()* is used for loading the per-calculated co-occurrence counts.

ii. Training the model with SGD. Initialize all model parameters in a new function *init_vectors()*, then, at each iteration, sample a non-zero element of the co-occurrence matrix and use its value to update the parameters of the model[2]. Run the training procedure over the entire matrix for 100 epochs, and plot the value of the cost function during training.

iii. Visualize the resulting word vectors, using the procedure *visualize()*, by projecting them on the plane defined by their two principal components. (As mentioned in the paper, create the final word vectors by summing the resulting word and context-word matrices $W+\widetilde{W}$).

iv. Read about Simlex-999. This is a dataset that contains word pairs and the similarity scores humans assigned to them. E.g., one pair of words is "bad" and "awful", and the similarity score is 8.42. The notebook contains code for loading the simlex-999 data. Do the word embeddings we calculated reflect human notion of similarity? Plot the simlex-similarity between each pair (w1, w2) and the cosine similarity of the corresponding word vectors. What is the Pearson correlation between the two? Why do we not get perfect correlation?

1 Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global Vectors for Word Representation. In *EMNLP* (Vol. 14, pp. 1532-43).

2 Set the weighting function as in the paper - Eqn. (9): $x_{max}=100, \alpha=0.75$.