TDA 231 Machine Learning 2017: Final Exam

Instructor: Devdatt Dubhashi

Due:10 AM, Room 6453, March 14, 2017

1. (15 points) Suppose we have a prediction problem in marine shipping where the target t corresponds to an angle measured in radians. A reasonable loss function in this case could be

$$\mathcal{L}(y,t) := 1 - \cos(y-t).$$

   (a) Suppose the true angle is $3\pi/4$. For which predicted values is the loss maximum and for which is it minimum? (Give also the corresponding loss.)

   Suppose we make predictions with a linear model

$$y = \mathbf{w}^T\mathbf{x} + b.$$

   (b) Derive gradients of the loss wth respect to both $\mathbf{w}$ and b.

   (c) Describe a gradient descent algorithm to train the model on a data set $(\mathbf{x}_i, t_i), i = 1 \cdots N$.

2. (15 points) Consider the following data set of an insurance company.:

customer	Age	Car	Premium
1	25	sports	L
2	20	vintage	H
3	25	sports	L
4	45	suv	H
5	20	sports	H
6	25	suv	H

   (a) Describe and fit a Bayes classifier using MLE and classify the new customer (23, sports) as L or H.

   (b) Describe and fit a Näive-Bayes classifier using MLE and classify the new customer (23, sports) as L or H.

   (c) What would happen if you get a new customer (23, truck)? How would you handle this?

3. (15 points) Recall how K–means++ initializes the centers for clustering. We initialize with $C_1 = \{\mathbf{x}\}$ where $\mathbf{x}$ is chosen uniformly at random from one of the input points. Then, for $2 \leq i \leq k$, $C_i := C_{i-1} \cup \{\mathbf{x}\}$ where $\mathbf{x}$ is chosen from the complement of C_{i-1} with probability

$$\nu(\mathbf{x}) = \frac{d^2(\mathbf{x}, C_{i-1})}{\sum_{\mathbf{y} \notin C_{i-1}} d^2(\mathbf{y}, C_{i-1})} \tag{1}$$

   Suppose we have N data points in dimension D.

   (a) What is the time complexity of this initialization step in terms of N, D and k?

   The rest of the problem is to alleviate the problem in (a) for Big Data using a MCMC approach. Consider the step of choosing the next center. Rather than considering all input points, we will develop a Markov Chain whose state space is the set of points in the complement of C_{i-1} and whose stationary distribution is given by (1).

(b) Give a proposal move.

(c) Give the Metropolis–Hastings acceptance rule tailored so that the stationary distribution is (1).

(d) Describe how to use the Markov Chain to give an alternative way to initialize the centers in K–Means ++.

(e) What is the running time of this initialization?

4. (15 points) Consider a neural network with N inputs, N outputs and a single hidden layer with K units. The activations are computed as follows:

$$\begin{aligned} \mathbf{z} &= \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} \\ \mathbf{h} &= \sigma(\mathbf{z}) \\ \mathbf{y} &= \mathbf{x} + \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)} \end{aligned}$$

Here the non–linear activation function is the sigmoid: $\sigma(x) = \frac{1}{1+e^{-x}}$. The cost will involve both $\mathbf{h}$ and $\mathbf{y}$:

$$\begin{aligned} \mathcal{E} &:= \mathcal{R} + \mathcal{S} \\ \mathcal{R} &:= \mathbf{r}^T\mathbf{h} \\ \mathcal{S} &:= \frac{1}{2}\|\mathbf{y} - \mathbf{s}\|^2 \end{aligned}$$

for given fixed vectors $\mathbf{r}$ and $\mathbf{s}$.

(a) Draw the computation graph showing connections in the network relating $\mathbf{x}, \mathbf{z}, \mathbf{h}, \mathbf{y}$ and $\mathcal{E}, \mathcal{R}, \mathcal{S}$.

(b) What is the total number of parameters in the model?

(c) Write down the gradients of the error $\mathcal{E}$ wth respect to all the parameters. Show an outline of your derivation.