UNIVERSITY OF OXFORD

Machine Learning
Hilary Term 2016
Week 6

Problem Sheet 4

In this problem sheet, the nature of questions will be substantially different from the previous ones; you should be well-versed with routine algebraic manipulations by now. The first problem focuses on a dimensionality reduction technique, known as the *Johnson-Lindenstrauss* Lemma.[1] For those of you who are more mathematically inclined, I suggest you try to prove the Lemma; we've provided a framework to guide you through the major steps. Others should simply focus on answering questions regarding the implications of the lemma. The second question looks at a simple three to four-layered neural network. We'll study the mathematical aspects of neural networks in the next two weeks, however, you should already be able to reason about the kind of questions posed here. The last part of this sheet requires you to read an article. There are a lot of technical details in the article which we don't expect you to understand fully. However, you should be able to grasp what is going on well enough to answer the questions we have framed in this sheet.

1 The Johnson-Lindenstrauss Lemma

(Courtesy: Rodrigo Mendoza-Smith) Note: This question looks scary, but it is quite straightforward and fun. Everyone is encouraged to solve the entire question; however, at the very least you should answer parts 8, 9, 10 (you may assume that the previous parts are true).

Suppose we have a dataset $S = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathbb{R}^n$ and n is very *very* large. Clearly, learning with the dataset S requires computation with vectors of dimension n, which can be expensive when n is so large. Hence, we would like to reduce the dimensionality of S while preserving geometric information about S. This is particularly useful for algorithms that learn from the geometry of the dataset such as clustering algorithms and nearest neighbour algorithms.

The Johnson-Lindenstrauss Lemma states that,

Lemma (Johnson-Lindenstrauss)**.** *For all $\varepsilon \in (0, 1)$ and any integer m, let k be a positive integer such that*

$$k \geq 4 \left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3} \right)^{-1} \log m. \tag{1}$$

Then, for all $S \subset \mathbb{R}^n$ with m points, there exists a function $f : \mathbb{R}^n \to \mathbb{R}^k$ such that,

$$(1 - \varepsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \varepsilon)\|u - v\|^2 \quad \forall u, v \in S$$

In this question we will prove Lemma 1.

[1] It is one of the few examples where lemmas have become more famous than the theorems they were first used to prove!

UNIVERSITY OF OXFORD

Machine Learning
Hilary Term 2016
Week 6

1. Let $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$ be i.i.d., $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = \frac{\mathbf{X}}{\|\mathbf{X}\|}$. Let $\Psi = [\mathbf{I}_k, \mathbf{0}_{k\times(n-k)}] \in \mathbb{R}^{k\times n}$, where $\mathbf{I}_k \in \mathbb{R}^{k\times k}$ is an identity matrix and $\mathbf{0}_{k\times n-k} \in \mathbb{R}^{k\times(n-k)}$ is a zero matrix. Show that,
$$\mathbb{E}\left[\|\Psi\mathbf{Y}\|^2\right] = \frac{k}{n}$$
Hint: Argue that $X_1^2\|\mathbf{X}\|^{-2}, \ldots, X_n^2\|\mathbf{X}\|^{-2}$ are all identically distributed (though not independent) and use linearity of expectation

2. We will need the following lemma:

   Lemma. *Let $k < n$. Then,*

   - *If $\beta < 1$,*
   $$\Pr\left[\|\Psi\mathbf{Y}\|^2 \leq \beta\frac{k}{n}\right] \leq \exp\left(\frac{k}{2}(1-\beta+\log\beta)\right). \tag{2}$$
   - *If $\beta > 1$,*
   $$\Pr\left[\|\Psi\mathbf{Y}\|^2 \geq \beta\frac{k}{n}\right] \leq \exp\left(\frac{k}{2}(1-\beta+\log\beta)\right). \tag{3}$$

   Suppose $\beta < 1$.

   (a) Show that for $t > 0$,
   $$\Pr\left[\|\Psi\mathbf{Y}\|^2 \leq \beta\frac{k}{n}\right] = \Pr\left[\exp\left\{t\left(k\beta\sum_{i=1}^{n}X_i^2 - n\sum_{i=1}^{k}X_i^2\right)\right\} \geq 1\right]$$

   (b) Use Markov's inequality[2] and use the result from the previous part to show that,
   $$\Pr\left[\|\Psi\mathbf{Y}\|^2 \leq \beta\frac{k}{n}\right] \leq \mathbb{E}\left[\exp\left\{t\left(k\beta\sum_{i=1}^{n}X_i^2 - n\sum_{i=1}^{k}X_i^2\right)\right\}\right] \tag{4}$$

   (c) If X, Y are independent random variables, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Use this fact, and the moment generating function of the Chi-Square distribution[3] to show that the right hand side of (4) is equal to,
   $$g(t) = (1-2tk\beta)^{-(n-k)/2}(1-2t(k\beta-n))^{-k/2}.$$

   (d) Argue that for the upper bound $g(t)$ to make sense we need $t \in (0, (2k\beta)^{-1})$.

   (e) Show that the minimiser of $g(t)$ subject to $t \in (0, (2k\beta)^{-1})$ is given by $t_0 = (1-\beta)(2\beta(n-k\beta))^{-1}$. *(Hint: argue that you can instead maximise $g(t)^{-2}$.)*

[2] `https://en.wikipedia.org/wiki/Markov%27s_inequality`
[3] `https://en.wikipedia.org/wiki/Moment-generating_function`

(f) Use the previous result to show that

$$\Pr\left[\|\Psi\mathbf{Y}\|^2 \leq \beta\frac{k}{n}\right] \leq \beta^{k/2}\left(1+\frac{k(1-\beta)}{n-k}\right)^{(n-k)/2}. \tag{5}$$

(g) Upper bound the right hand side of (5) to recover (2). *Hint: If* $t, x > 0$, $e^t \geq \left(1+\frac{t}{x}\right)^x$.

(h) Follow the same approach to prove (3).

(i) Based on Lemma 2, is it likely that $\|\Psi\mathbf{Y}\|^2$ will be far away from its mean?

3. (Let's go back to proving Lemma 1) Suppose that $n > k$ (Otherwise we are done). Let $\Phi \in \mathbb{R}^{k\times n}$ denote a linear transformation that projects $\mathbb{R}^n$ into a *random* k dimensional subspace of $\mathbb{R}^n$. Use any of the previous results to show that if $\mathbf{u}, \mathbf{v} \in S$,

$$\mathbb{E}\left[\|\Phi(\mathbf{u}-\mathbf{v})\|^2\right] = \frac{k}{n}\|\mathbf{u}-\mathbf{v}\|^2$$

Argue that Lemma 2 can be used in this case.

4. Let $\varepsilon \in (0,1)$. Use Lemma 2 to show that for any $\mathbf{u}, \mathbf{v} \in S$ with $\mathbf{u} \neq \mathbf{v}$

$$\Pr\left[\|\Phi(\mathbf{u}-\mathbf{v})\|^2 \leq (1-\varepsilon)\frac{k}{n}\|\mathbf{u}-\mathbf{v}\|^2\right] \leq \exp\left(-k\frac{\varepsilon^2}{4}\right) \tag{6}$$

Hint: $\log(1-x) \leq -x - x^2/2$ *for all* $0 \leq x < 1$.

5. Let k satisfy (1). Use (6) to show that

$$\Pr\left[\|\Phi(\mathbf{u}-\mathbf{v})\|^2 \leq (1-\varepsilon)\frac{k}{n}\|\mathbf{u}-\mathbf{v}\|^2\right] \leq \frac{1}{m^2}. \tag{7}$$

6. (Optional) Follow the same reasoning to show that

$$\Pr\left[\|\Phi(\mathbf{u}-\mathbf{v})\|^2 \geq (1+\varepsilon)\frac{k}{n}\|\mathbf{u}-\mathbf{v}\|^2\right] \leq \frac{1}{m^2}. \tag{8}$$

Hint: Use Lemma 2 and the fact that $\log(1+x) \leq x - x^2/2 + x^3/3$.

7. Use (7) and (8) to propose an $f : \mathbb{R}^n \to \mathbb{R}^k$ such that for $\mathbf{u}, \mathbf{v} \in S$ with $\mathbf{u} \neq \mathbf{v}$

$$\Pr\left[\frac{\|f(\mathbf{v})-f(\mathbf{u})\|^2}{\|\mathbf{u}-\mathbf{v}\|^2} \notin [(1-\varepsilon),(1+\varepsilon)]\right] \leq \frac{2}{m^2}. \tag{9}$$

8. Use a union bound[4] to show that

$$\Pr\left[\exists \mathbf{u}, \mathbf{v} \in S : \mathbf{u} \neq \mathbf{v} \text{ and } \frac{\|f(\mathbf{v})-f(\mathbf{u})\|^2}{\|\mathbf{u}-\mathbf{v}\|^2} \notin [(1-\varepsilon),(1+\varepsilon)]\right] \leq 1 - \frac{1}{m}.$$

Hint: Rewrite the event $\{\exists \mathbf{u}, \mathbf{v} \in S : \mathbf{u} \neq \mathbf{v}$ *and* $\frac{\|f(\mathbf{v})-f(\mathbf{u})\|^2}{\|\mathbf{u}-\mathbf{v}\|^2} \notin [(1-\varepsilon),(1+\varepsilon)]\}$ *as the union of* $\binom{m}{2}$ *events.*

[4] https://en.wikipedia.org/wiki/Boole%27s_inequality

9. Does the previous result imply Lemma 1? *Hint: yes.*

10. How would you use Lemma 1 for the k Nearest Neighbours? Does it improve the computational complexity?

2 Digit Classification Using Neural Networks

In this problem, we will consider a neural network to classify handwritten digits. You have already seen this dataset in your practical last week. However, we consider a slightly different network and a different way to encode the targets (and the outputs of the network).

The inputs are vectors of length 1024 (obtained from 32×32 grey pixel images). Rather than output a class, the network will output a vector $\hat{\mathbf{y}} \in \mathbb{R}^{10}$. The target will be represented as a one-hot encoding, *e.g.,* if the label is 3, we will use the vector $(0, 0, 1, 0, 0, 0, 0, 0, 0, 0)^T$ (by convention 0 will be the last component, not the first).

The data fed into the neural network consists of $\langle(\mathbf{x}_i, \mathbf{y}_i)\rangle$, where $\mathbf{x}_i \in \mathbb{R}^{1024}$ and $\mathbf{y}_i \in \{0, 1\}^{10}$ is the one-hot encoding of the digit. The output of the neural network is also a vector $\hat{\mathbf{y}} \in \mathbb{R}^{10}$. We consider minimizing the squared loss objective (with respect to the parameters of the neural network):

$$\mathrm{L}(\mathbf{w}) = \sum_{i=1}^{m} (\hat{\mathbf{y}}_i - \mathbf{y}_i)^2$$

The neural network to be used is shown in Figure 1. There are three layers, one input layer, one *hidden* layer and one output layer. We will study in class how the parameters can be adjusted by using backpropagation. However, here you are asked to think about the following questions:

1. To increase efficiency, your colleague suggests that instead of using the one-hot encoding, use the standard binary encoding; so 0 is 0000, 1 is 0001, and so on, 9 is 1001. Thus, your output layer only needs to have four neurons rather than 10. Perhaps you could also reduce the number of neurons in the middle layer. What do you think about this suggestion?

2. Show that if you have a well-trained neural network that has high accuracy with the *one-hot* encoding, you can design a network that uses binary encoding and achieves the same error. *(Hint: You may need to add an additional layer.)*

3. What do you think would happen if you tried to train the neural network you suggested in the previous part directly (rather than adding the last layer by design)?

3 Reducing Multiclass to Binary

You should read the following article:

- **Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers.** E. L. Allwein, Robert E. Schapire, Y. Singer. Journal of Machine Learning Research (2000).

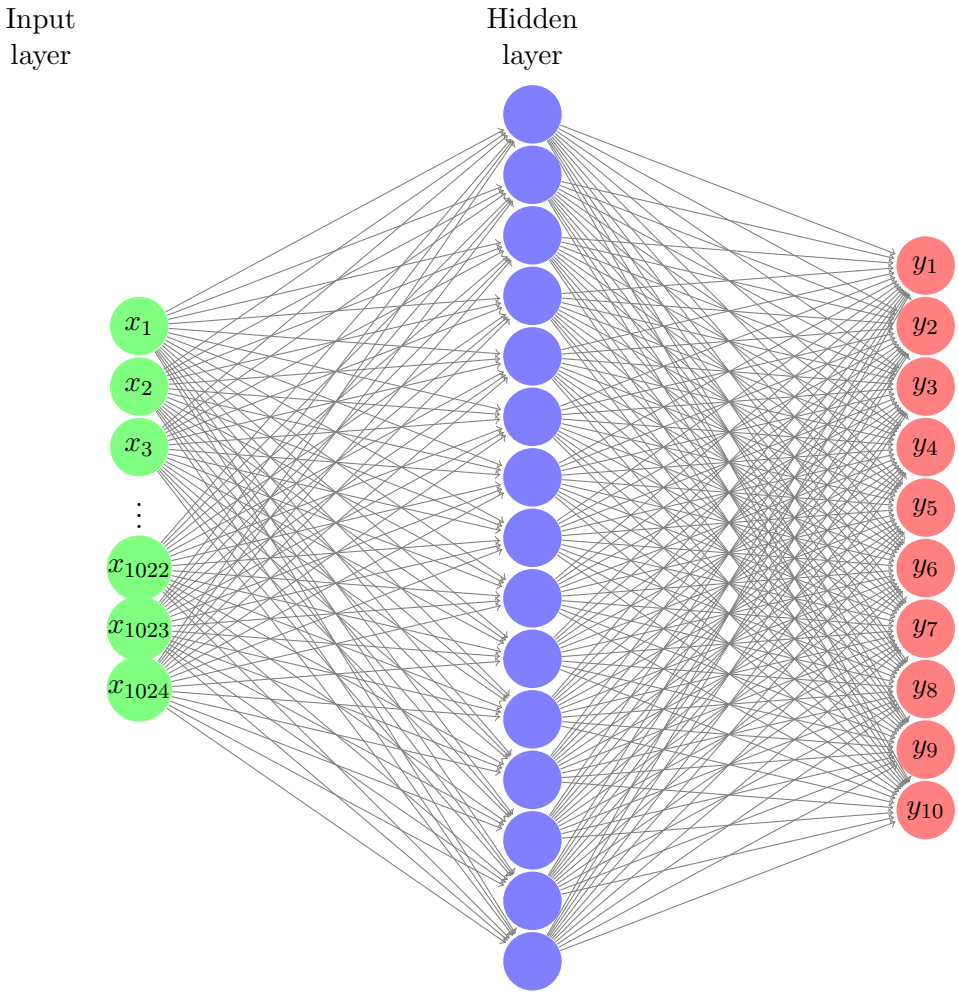

Figure 1: Three-layer neural network for handwritten digit classification.

The article is linked on the class-website under the section "Resources". There are also slides based on the article available at: `www.cs.princeton.edu/~schapire/talks/ecoc-icml10.pdf`.

Reading Guidelines: You are not expected to understand everything in the article. You may skip Sections 4 and 5 from the paper entirely, as they are concerned with proving theoretical bounds on approaches studied in this paper; of course, if you so wish, you are welcome to read them. We have studied some of the classifiers they mention in class, you may skip the description of classifiers that look unfamiliar. Suggest answers to the following questions:

- What are the main differences between the error correcting code approach (ECOC) proposed by Dietterich and Bakiri and the approaches suggested in this paper?

- Why would it be better to allow the coding matrix to contain 0 entries, rather than just ± 1?
- What do the authors propose as an explanation for why loss based decoding is better than Hamming decoding?
- For what values of k, do you expect the coding approach to offer a genuine advantage over the one-vs-one or one-vs-all methods in terms of computational complexity? Do you think that the experiments in the paper give evidence that this is indeed the case? *(While you should not hesitate to scientifically criticize any aspects of the paper, bear in mind that the technology used in this paper is from 1999.)*