Machine Learning: Homework Assignment 1

E4525 Spring 2019,

IEOR, Columbia University

Due: February 1st, 2019

1. **ML Paper review** Skim through the papers l [Tiwari et al., 2016] and [Esteva et al., 2017]. You don't need to read them carefully, or understand them in any detail.
   Answer the following questions:

   (a) For [Tiwari et al., 2016]'s paper:

      i. What are their inputs, what is their source of data?
         MRI scans of patient brains
      ii. What is the medical problem they are trying to identify?
         Recurrence of a brain tumor after treatment
      iii. Into how many classes do they classify their images?
         Two: Recurrence or no recurrence
      iv. How many data samples do they use?
         less than 100.
      v. How do they evaluate the performance of their algorithm?
         They compare diagnosis with two trained Radiologists

   (b) For [Esteva et al., 2017]'s paper:

      i. What are their inputs, what is their source of data?
         Clinical images of skin lesions
      ii. What is the medical problem they are trying to identify?
         A variety of different skin cancers.
      iii. How many disease classes do they train their classifier to recognize (only consider the finer level of their classification)?
         The dataset contains $\approx 2,000$ disease classes but the CNN is trained using 757 disease classes. Later to test accuracy they group diseases in 9 categories. [757 is the correct answer, partial credit for 9]
      iv. How many data samples do they use?
         $\approx 100,000$
      v. How do they evaluate the performance of their algorithm?
         They compare to the predictions of 21 board-certified dermatologists

2. **Scalar data types**: Classify each one of this variables into one of

   - Categorical
   - Ordinal
   - Interval
   - Ratio

   (a) Number of patients in a hospital
   Ratio

   (b) Bronze, Silver, Gold medals as awarded in the Olympics
   Ordinal

   (c) Student Id number
   Categorical

   (d) Film classification into: Comedy, Drama, etc.
   Categorical

   (e) Distance in meters as measured from the surface of the earth.
   Ratio

   (f) A homework assignment grade in a 0 to 100 scale
   Interval, it does not really make sense to compute ratios of grades.

   (g) A course grade in a E to A+ scale.
   Ordinal, the spacing between grade letter grades is not uniform.

   (h) email address
   Categorical

   (i) An angle between 0 and 360 degrees.
   Interval, the ratio of two angles is meaningless.

3. **Vector representation of Binary Variables**

   Let's assume two binary statements X and Y that can be either true of false.

   We jointly observe N samples of X and Y and record the results on the following N dimensional vectors

   - $Z^X = \{z_i^X\}$, where $i = 1, \ldots N$. $z_i^X = 1$ if statement X was true on sample i, zero otherwise.
   - $Z^Y = \{z_i^Y\}$, where $i = 1, \ldots N$. $z_i^Y = 1$ if statement Y was true on sample i, zero otherwise.

   In what follows the operator $A * B$ denotes **element wise** multiplication $(A * B)_i = A_i B_i$, the `sum` of a vector is $\text{sum}(Z) = \sum_{i=1}^{N} Z_i$, and the dot product of two vectors is $A^T \cdot B = \text{sum}(A * B) = \sum_{i=1}^{N} A_i B_i$

- Write a mathematical expression (in terms of z_i^X) for the total number of samples in which the statement X was true

$$\hat{N}^X = \sum_{i=1}^{N} z_i^X \tag{1}$$

- Write a mathematical expression for the fraction of samples in which statement X was true

$$\bar{z}^X = \frac{1}{N} \sum_{i=1}^{N} z_i^X \tag{2}$$

- What is the interpretation of the following vector expression?

$$Z^{XY} = Z^X * Z^Y \tag{3}$$

  Z^{XY} is a variable that is 1 for samples where both X and Y are true at the same time.

- What is the interpretation of the ordinary dot product of the vectors Z^X and Z^Y?

  It it the number of times in which both X and Y are true at the same time

- Write a vector expression for the proportion of samples in with X was true but Y was false.

$$\hat{p}^{X\bar{Y}} = \frac{1}{N} Z^X \cdot (1 - Z^Y) \tag{4}$$

- Write a vector form expression to compute the number of times that either X, Y or both were true.

$$\hat{N}^{X \vee Y} = \text{sum}(Z^X + Z^Y - Z^X * Z^Y) \tag{5}$$

  There are a few other variations.

- Write a vector form expression to compute the number of times than only one of X or Y was true, but not both.

$$\hat{N}^{X \bar{\oplus} Y} = (Z^X - Z^Y)^T (Z^X - Z^Y) \tag{6}$$

4. **Matrix and Index Notation**:

   Let's assume a D-dimensional regression model

$$y = x_1\theta_1 + x_2\theta_2 + \cdots + x_D\theta_D + \epsilon = \sum_{d=1}^{D} x_d\theta_d + \epsilon \tag{7}$$

where ϵ is some random noise term with zero mean.

Given

- a matrix of observations $X = \{x_{i,d}\}$ where $i = 1, \dots, N$ runs through N observations, and $d = 1, \dots, D$ runs through D variables.
- a vector of outcomes $Y = \{y_i\}$
- a vector of noise terms $\mathcal{E} = \{\epsilon_i\}$
- a vector of parameters $\Theta = \{\theta_d\}$.

(a) Write in matrix notation (using X, Y, Θ and $\mathcal{E}$) an equation relating the outcome vector Y to the observations X and the noise $\mathcal{E}$

$$Y = X\Theta + \mathcal{E} \tag{8}$$

(b) Write a matrix expression for the average square errors (the average of the square of ϵ_i) in terms of X,Y and Θ

$$E = \frac{1}{N}(Y - X\theta)^T(Y - X\theta) \tag{9}$$

(c) Write an **explicit** expression for the average square error E in terms of the indexed variables ($x_{i,d}$, θ_d, y_i). Be explicit with the summation indexes.

$$E = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \sum_{d=1}^{D} x_{i,d}\theta_d\right)^2 \tag{10}$$

(d) To minimize the square error E relative to the parameters θ_d we must solve the first order conditions

$$\frac{\partial E}{\partial \theta_d} = 0 \tag{11}$$

Find explicit expressions for $\frac{\partial E}{\partial \theta_d}$ and write the equations 11 in terms of the indexed variables $x_{i,d}$, etc.

$$0 = \frac{\partial E}{\partial \theta_d} = \frac{2}{N}\sum_i\left(\sum_{d'=1}^{D} x_{i,d'}\theta_{d'} - y_i\right)x_{i,d} \tag{12}$$

When grading pay special attention that the dummy index d' and the free index d are not mixed-up in the student's answer.

(e) Translate those equations into a matrix equation for Θ

$$(X^T X)\Theta = X^T Y \tag{13}$$

5. **Matrix and Index Notation II**:

Let's assume a D-dimensional regression model for a K-dimensional outcome vector

$$y_1 = \sum_{d=1}^{D} x_d w_{1,d} + \epsilon_1$$
$$\vdots \tag{14}$$
$$y_k = \sum_{d=1}^{D} x_d w_{k,d} + \epsilon_k$$
$$\vdots \tag{15}$$
$$y_K = \sum_{d=1}^{D} x_d w_{K,d} + \epsilon_K \tag{16}$$

where ϵ_k, $k = 1, \ldots, K$ is some random noise term with zero mean. ϵ_k is independent from $\epsilon_{k'}$ when $k \neq k'$.

Given

- a matrix of observations $X = \{x_{i,d}\}$ where $i = 1, \ldots, N$ runs through N observations, and $d = 1, \ldots, D$ runs through D variables.
- a matrix of outcomes $Y = \{y_{i,k}\}$ where $k = 1, \ldots, K$.
- a matrix of noise terms $\mathcal{E} = \{\epsilon_{i,k}\}$, were $\epsilon_{i,k} \sim \mathcal{N}(0, \sigma^2)$.
- a matrix of parameters $W = \{w_{k,d}\}$.

(a) Write in matrix notation (using X, Y, W, and $\mathcal{E}$) an equation relating the outcome vector Y to the observations X and the noise $\mathcal{E}$

$$Y = XW^T + \mathcal{E} \tag{17}$$

(b) Write a matrix expression for the average square errors (the average over the observations i of the sum over k of $\epsilon_{i,k}$) in terms of X,Y and W.

[Hint] You may need to use the matrix trace function $\mathrm{tr}(A) = \sum_i A_{i,i}$.

$$E = \frac{1}{N} \mathrm{tr}\left[(Y - XW^T)^T (Y - XW^T)\right] \tag{18}$$

[Note, because tr[AB] = tr[BA] , the expression

$$E = \frac{1}{N}\mathrm{tr}\left[(Y - XW^T)(Y - XW^T)^T\right] \tag{19}$$

is equivalent and a valid answer too.]

(c) Write an **explicit** expression for the average square error E in terms of the indexed variables ($x_{i,d}$, $w_{k,d}$, $y_{i,k}$).

$$E = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N}\left(y_{i,k} - \sum_{d=1}^{D} x_{i,d}w_{k,d}\right)^2 \tag{20}$$

Be explicit with the summation indexes.

(d) To minimize the square error E relative to the parameters $w_{k,d}$ we must solve the first order conditions

$$\frac{\partial E}{\partial w_{k,d}} = 0 \tag{21}$$

Find explicit expressions for $\frac{\partial E}{\partial w_{k,d}}$ and write the equations 21 in terms of the indexed variables $x_{i,d}$, etc.

$$0 = \frac{\partial E}{\partial w_{k,d}} = -\frac{2}{N}\sum_{i=1}^{N}\left(y_{i,k} - \sum_{d'=1}^{D} x_{i,d'}w_{k,d'}\right)x_{i,d} \tag{22}$$

When grading pay special attention that the dummy index d' and the free index d are not mixed-up in the student's answer.

(e) Translate the equations derived in exercise 5d into a matrix equation for W

$$X^TY = (X^TX)W^T \tag{23}$$

References

[Esteva et al., 2017] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118. `https://cs.stanford.edu/people/esteva/nature/#!`

[Tiwari et al., 2016] Tiwari, P., Prasanna, P., Wolansky, L., Pinho, M., Cohen, M., Nayate, A., Gupta, A., Singh, G., Hattanpaa, K., Sloan, A., Rogers, L., and Madabhushi, A. (2016). Computer-extracted texture features to distinguish cerebral radionecrosis from recurrent brain tumors on multiparametric mri: A feasibility study. *American Journal of Neuroradiology*. https://doi.org/10.3174/ajnr.A4931.