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?
      ii. What is the medical problem they are trying to identify?
      iii. Into how many classes do they classify their images?
      iv. How many data samples do they use?
      v. How do they evaluate the performance of their algorithm?
   (b) For [Esteva et al., 2017]'s paper:
      i. What are their inputs, what is their source of data?
      ii. What is the medical problem they are trying to identify?
      iii. How many disease classes do they train their classifier to recognize (only consider the finer level of their classification)?
      iv. How many data samples do they use?
      v. How do they evaluate the performance of their algorithm?

2. **Scalar data types**: Classify each one of this variables into one of
   - Categorical
   - Ordinal
   - Interval
   - Ratio

   (a) Number of patients in a hospital
   (b) Bronze, Silver, Gold medals as awarded in the Olympics
   (c) Student Id number
   (d) Film classification into: Comedy, Drama, etc.
   (e) Distance in meters as measured from the surface of the earth.

(f) A homework assignment grade in a 0 to 100 scale

(g) A course grade in a E to A+ scale.

(h) email address

(i) An angle between 0 and 360 degrees.

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 $i = 1, \dots N$. $z_i^X = 1$ if statement X was true on sample i, zero otherwise.
   - $Z^Y = \{z_i^Y\}$, where $i = 1, \dots N$. $z_i^Y = 1$ 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
   - Write a mathematical expression for the fraction of samples in which statement X was true
   - What is the interpretation of the following vector expression?

     $$Z^{XY} = Z^X * Z^Y \tag{1}$$

   - What is the interpretation of the ordinary dot product of the vectors Z^X and Z^Y?
   - Write a vector expression for the proportion of samples in with X was true but Y was false.
   - Write a vector form expression to compute the number of times that either X, Y or both were true.
   - Write a vector form expression to compute the number of times than only one of X or Y was true, but not both.

4. **Matrix and Index Notation**:

   Let's assume a D-dimensional regression model

   $$y = x_1\theta_1 + x_2\theta_2 + \cdots + x_D\theta_D + \epsilon = \sum_{d=1}^{D} x_d\theta_d + \epsilon \tag{2}$$

   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 $\mathcal{E}$

(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

(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.

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

$$\frac{\partial E}{\partial \theta_d} = 0 \tag{3}$$

Find explicit expressions for $\frac{\partial E}{\partial \theta_d}$ and write the equations 3 in terms of the indexed variables $x_{i,d}$, etc.

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

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

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

$$\begin{aligned} y_1 &= \sum_{d=1}^{D} x_d w_{1,d} + \epsilon_1 \\ &\vdots \\ y_k &= \sum_{d=1}^{D} x_d w_{k,d} + \epsilon_k \\ &\vdots \\ y_K &= \sum_{d=1}^{D} x_d w_{K,d} + \epsilon_K \end{aligned} \tag{4--6}$$

where ϵ_k, $k = 1, \dots, K$ 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 $i = 1, \dots, N$ runs through N observations, and $d = 1, \dots, D$ 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 $\mathcal{E}$

(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 $\mathrm{tr}(A) = \sum_i A_{i,i}$.

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

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

$$\frac{\partial E}{\partial w_{k,d}} = 0 \tag{7}$$

Find explicit expressions for $\frac{\partial E}{\partial w_{k,d}}$ and write the equations 7 in terms of the indexed variables $x_{i,d}$, etc.

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

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`.