CSE 250B: Machine learning **Winter 2018**

Homework 3 — Generative models, exponential families, linear algebra

By the due date (midnight of Wednesday Jan 31), upload the PDF of your `typewritten` solutions to `gradescope`. Problems 6–19 are intended as a review of basic linear algebra.

1. *Handwritten digit recognition using a Gaussian generative model.* Recall that you can obtain the MNIST data from `http://yann.lecun.com/exdb/mnist/index.html`. In this problem, you will build a classifier for this data by modeling each class as a multivariate (784-dimensional) Gaussian.

   (a) Split the training set into two pieces – a training set of size 50000, and a separate *validation set* of size 10000. Also load in the test data.

   (b) Now fit a Gaussian generative model to the training data of 50000 points:

   - Determine the class probabilities: what fraction π_0 of the training points are digit 0, for instance? Call these values $\pi_0, \ldots, \pi_9$.
   - Fit a Gaussian to each digit, by finding the mean and the covariance of the corresponding data points. Let the Gaussian for the jth digit be $P_j = N(\mu_j, \Sigma_j)$.

   Using these two pieces of information, you can classify new images x using Bayes' rule: simply pick the digit j for which $\pi_j P_j(x)$ is largest.

   (c) One last step is needed: it is important to smooth the covariance matrices, and the usual way to do this is to add in cI, where c is some constant and I is the identity matrix. What value of c is right? Use the validation set to help you choose.

   (d) Turn in the following details:

   - A (very) short description of your procedure for choosing c.
   - A graph showing the validation error for all the values of c you tried.
   - Your final choice of c.
   - Overall error rate on the MNIST test set.

2. Compute the entropy of each of the following random variables X.

   (a) X is a coin of bias $2/3$.

   (b) X is the outcome of rolling a fair die.

   (c) $X = (X_1, \ldots, X_{10})$, where the X_i are independent and are each coins of bias $1/2$.

   (d) $X = (X_1, \ldots, X_{10})$ where X_1 is a coin of bias $1/2$ and $X_i = X_1 + i$.

3. *Entropy of continuous distributions.* As defined in class, a distribution P over a *discrete* set S has entropy

$$H(P) \;=\; \sum_{x \in S} P(x) \log \frac{1}{P(x)} \;=\; \mathbb{E}_{X \sim P}\left[\log \frac{1}{P(X)}\right],$$

   where $X \sim P$ means that random variable X is distributed according to P. The second term shows how this notion can be generalized to the continuous setting: for a density p, we define the *differential entropy* as

$$h(p) = \int p(x) \log \frac{1}{p(x)}\, dx.$$

(a) Show that the differential entropy of the one-dimensional Gaussian with mean μ and variance σ^2 is $\frac{1}{2}\log(2\pi\sigma^2 e)$.

(b) What is the differential entropy of the uniform distribution over $[a, b]$? Why is this suspicious when $b - a$ is small?

4. You come across a mysterious new language and find, after looking at some examples of words, that

   - Each word is composed of characters 'a' through 'z'
   - Each word has a length of at most 15
   - The average word length is 4
   - 80% of words end with a vowel
   - In 90% of words, every consonant is followed by a vowel
   - The first letter of a word is 'z' 50% of the time

   You decide to use the maximum entropy approach to come up with a simple distribution for the words in this language. What is the set S, and what features T would you choose? What is the functional form of the maximum entropy solution?

5. Consider the family of distributions on the positive reals $[0, \infty)$ parametrized by $\lambda > 0$, with density $p(x) = \lambda e^{-\lambda x}$. Show that this is an exponential family by exhibiting a suitable (S, π, T). What is the natural parameter space Θ?

6. Find the unit vector in the same direction as $x = (1, 2, 3)$.

7. Find all unit vectors in $\mathbb{R}^2$ that are orthogonal to $(1, 1)$.

8. How would you describe the set of all points $x \in \mathbb{R}^d$ with $x \cdot x = 25$?

9. The function $f(x) = 2x_1 - x_2 + 6x_3$ can be written as $w \cdot x$ for $x \in \mathbb{R}^3$. What is w?

10. For a certain pair of matrices A, B, the product AB has dimension 10×20. If A has 30 columns, what are the dimensions of A and B?

11. We have n data points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$ and we store them in a matrix X, one point per row.

    (a) What is the dimension of X?

    (b) What is the dimension of XX^T?

    (c) What is the (i, j) entry of XX^T, simply?

12. Vector x has length 10. What is $x^T x x^T x x^T x$?

13. For $x = (1, 3, 5)$ compute $x^T x$ and xx^T.

14. Vectors $x, y \in \mathbb{R}^d$ both have length 2. If $x^T y = 2$, what is the angle between x and y?

15. The quadratic function $f : \mathbb{R}^3 \to \mathbb{R}$ given by

$$f(x) = 3x_1^2 + 2x_1x_2 - 4x_1x_3 + 6x_3^2$$

    can be written in the form $x^T M x$ for some **symmetric** matrix M. What is M?

16. Which of the following matrices is necessarily symmetric?

(a) AA^T for arbitrary matrix A.
(b) $A^T A$ for arbitrary matrix A.
(c) $A + A^T$ for arbitrary square matrix A.
(d) $A - A^T$ for arbitrary square matrix A.

17. Let $A = \text{diag}(1, 2, 3, 4, 5, 6, 7, 8)$.

    (a) What is $|A|$?
    (b) What is A^{-1}?

18. Vectors $u_1, \ldots, u_d \in \mathbb{R}^d$ all have unit length and are orthogonal to each other. Let U be the $d \times d$ matrix whose rows are the u_i.

    (a) What is UU^T?
    (b) What is U^{-1}?

19. Matrix $A = \begin{pmatrix} 1 & 2 \\ 3 & z \end{pmatrix}$ is singular. What is z?