

ENGR 108 Extra Practice Problems

A3.27 *A p th moment deviation bound.* The Chebyshev inequality states that the number k of elements of an n -vector x satisfying $|x_i| \geq a$ must satisfy $\frac{k}{n} \leq \frac{\|x\|_p^2}{a^2 n}$. We can define p -norms, which are distinct from Euclidean norms, for $1 \leq p < \infty$ by

$$\|x\|_p = \sqrt[p]{|x_1|^p + |x_2|^p + \cdots + |x_n|^p},$$

so $\|x\|_p^p = |x_1|^p + \cdots + |x_n|^p$. In analogy with **rms**, let $\bar{x}_n = \frac{1}{n}(x_1 + \cdots + x_n) = \frac{1}{n}\mathbf{1}^T x$ denote the mean of x , and define the “root-mean- p th” value

$$\mathbf{rmp}(x) = \sqrt[p]{\frac{1}{n}(|x_1 - \bar{x}_n|^p + \cdots + |x_n - \bar{x}_n|^p)} = \frac{1}{n^{1/p}} \|x - (\mathbf{1}^T x/n)\mathbf{1}\|_p.$$

- (a) Using a similar argument to that for the Chebyshev inequality, verify the following *p th moment bound*: if k is the number of elements of the n -vector x satisfying $|x_i| \geq a$, then

$$\frac{k}{n} \leq \frac{\|x\|_p^p}{na^p}.$$

- (b) Verify the following *mean p th moment bound*: if k is the number of elements of the n -vector x satisfying $|x_i - \frac{1}{n}\mathbf{1}^T x| \geq a$, then

$$\frac{k}{n} \leq \frac{\mathbf{rmp}(x)^p}{a^p}.$$

- (c) According to the mean p th moment bound, what is the largest percentage of entries in a vector x that can be farther than $2\mathbf{rmp}(x)$ units away from $\mathbf{avg}(x)$ for $p = 2$?
- (d) According to the mean p th moment bound, what is the largest percentage of entries in a vector x that can be farther than $2\mathbf{rmp}(x)$ units away from $\mathbf{avg}(x)$ for $p = 4$?
- (e) According to the mean p th moment bound, what is the largest percentage of entries in a vector x that can be farther than $2\mathbf{rmp}(x)$ units away from $\mathbf{avg}(x)$ for $p = 8$?
- 5.8 *Early termination of Gram–Schmidt algorithm.* When the Gram–Schmidt algorithm is run on a particular list of 10 15-vectors, it terminates in iteration 5 (since $\tilde{q}_5 = 0$). Which of the following must be true?
- (a) a_2, a_3, a_4 are linearly independent.
- (b) a_1, a_2, a_5 are linearly dependent.
- (c) a_1, a_2, a_3, a_4, a_5 are linearly dependent.
- (d) a_4 is nonzero.

A6.12 *Running average matrix.* Suppose the T -vector x represents a time series, with x_t its value in period t , for $t = 1, \dots, T$. The *running average* of x is another vector y , with

$$y_t = \frac{1}{t} (x_1 + \cdots + x_t), \quad t = 1, \dots, T.$$

(For example, $y_1 = x_1$ and $y_T = \mathbf{avg}(x)$.)

- (a) Find the $T \times T$ matrix R for which $y = Rx$. (We can call R the *running sum matrix*.) (To describe R you can say what R_{ij} is, for $i, j = 1, \dots, T$.)
- (b) Which of the following attributes does R have? Symmetric, diagonal, lower triangular, upper triangular. (You can choose more than one.)

- (c) Give a brief interpretation of your answer to part (b) that does not refer to matrices, but only to properties of forming the running average.

A8.4 Formulate the following problem as a set of linear equations in the form $Ax = b$. Give the numerical values of the elements of the 4×4 matrix A .

Find a polynomial $p(t) = x_1 + x_2t + x_3t^2 + x_4t^3$ that satisfies the four conditions

$$\int_0^1 p(t)dt = b_1, \quad \int_0^1 tp(t)dt = b_2, \quad \int_0^1 t^2p(t)dt = b_3, \quad \int_0^1 t^3p(t)dt = b_4.$$

The numbers b_1, b_2, b_3, b_4 on the right-hand sides are given.

A10.4 *Student-course matrix.* The Stanford registrar has the complete list of courses taken by each graduating student over several graduating classes. This data is represented by an $m \times n$ matrix C , with $C_{ij} = 1$ if student i took class j , and $C_{ij} = 0$ otherwise, for $i = 1, \dots, m$ and $j = 1, \dots, n$. (Thus, there are m students in the data set, and n different courses. For simplicity, we ignore the possibility that in some circumstances a student can take a course multiple times.)

Answer each of the questions below in English, with no equations, references to matrices or vectors, and so on. (You can refer to student i and course j , though.)

- (a) What is $(C^T C)_{kl}$?
 - (b) What is $(CC^T)_{rs}$?
 - (c) What is $(C^T \mathbf{1})_p$?
 - (d) Suppose you cluster the columns of C^T using k -means, with $k = 50$ (say). What do you think the results might look like? (Your response can be a bit vague, but not more than one or two sentences.)
 - (e) (Continuation of part (d).) Suppose z_1 is the cluster representative for group 1. What does $(z_1)_{143} = 0.01$ mean?
- 10.23 Suppose A is an $n \times n$ matrix that satisfies $A^2 = 0$. Does this imply that $A = 0$? (This is the case when $n = 1$.) If this is (always) true, explain why. If it is not, give a specific counterexample, *i.e.*, a matrix A that is nonzero but satisfies $A^2 = 0$.
- 11.3 *Matrix cancellation.* Suppose the scalars a , x , and y satisfy $ax = ay$. When $a \neq 0$ we can conclude that $x = y$; that is, we can cancel the a on the left of the equation. In this exercise we explore the matrix analog of cancellation, specifically, what properties of A are needed to conclude $X = Y$ from $AX = AY$, for matrices A , X , and Y ?
- (a) Give an example showing that $A \neq 0$ is not enough to conclude that $X = Y$.
 - (b) Show that if A is left-invertible, we can conclude from $AX = AY$ that $X = Y$.
 - (c) Show that if A is not left-invertible, there are matrices X and Y with $X \neq Y$, and $AX = AY$.

Remark. Parts (b) and (c) show that you can cancel a matrix on the left when, and only when, the matrix is left-invertible.

12.8 *Least squares and QR factorization.* Suppose A is an $m \times n$ matrix with linearly independent columns and QR factorization $A = QR$, and b is an m -vector. The vector $A\hat{x}$ is the linear combination of the columns of A that is closest to the vector b , *i.e.*, it is the projection of b onto the set of linear combinations of the columns of A .

- (a) Show that $A\hat{x} = QQ^T b$. (The matrix QQ^T is called the *projection matrix*.)

- (b) Show that $\|A\hat{x} - b\|^2 = \|b\|^2 - \|Q^T b\|^2$. (This is the square of the distance between b and the closest linear combination of the columns of A .)

13.9 *Conclusions from 5-fold cross-validation.* You have developed a regression model for predicting a scalar outcome y from a feature vector x of dimension 20, using a collection of $N = 600$ data points. The mean of the outcome variable y across the given data is 1.85, and its standard deviation is 0.32. After running 5-fold cross-validation we get the following RMS test errors (based on forming a model based on the data excluding fold i , and testing it on fold i).

Fold excluded	RMS test error
1	0.13
2	0.11
3	0.09
4	0.11
5	0.14

A13. How would you expect your model to do on new, unseen (but similar) data? Respond briefly and justify your response.

- (a) A co-worker observes that the regression model parameters found in the 5 different folds are quite close, but not the same. He says that for the production system, you should use the regression model parameters found when you excluded fold 3 from the original data, since it achieved the best RMS test error. Comment briefly.

A13.7 A daily time series of demand for some product, denoted y_t for day t , has been standardized based on its historical average and standard deviation, so $y_t = 0$ means the demand on day t has its historical average value, and $y_t = -1$ means the demand is one standard deviation below its historical average. When y_t is positive, say bigger than 0.3 or so, we say it is high demand day, and when y_t is negative, say smaller than -0.3 or so, we say it is a low demand day.

A data scientist develops a regression model $\hat{y}_{t+1} = x_t^T \beta + v$ that predicts tomorrow's demand y_{t+1} from the following two features: $(x_t)_1 = y_t$ is today's demand, and $(x_t)_2 = 1$ if tomorrow is a weekday, and $(x_t)_2 = 0$ if tomorrow is a weekend or official holiday. She finds that the model makes reasonably good predictions on test data, with RMS error around 0.1. Her model coefficients are $\beta = (0.8, -0.2)$, $v = 0.1$. Based on her model, which of the following statements are reasonable?

- (a) High demand days are typically followed by high demand days.
- (b) Low demand days are typically followed by high demand days.
- (c) Demand is typically much higher on weekends and holidays, compared to weekdays.
- (d) Demand typically decreases a bit on weekdays, compared to weekends and holidays.

A14.4 *Modifying a classifier.* A co-worker develops a classifier of the form $\hat{y} = \text{sign}(x^T \beta + v)$, with $v < 0$, where the n -vector x is the feature vector, and the n -vector β and scalar v are the classifier parameters. The classifier is evaluated on a given test data set. The false positive rate is the fraction of the test data points with $y = -1$ for which $\hat{y} = 1$. (We will assume there is at least one data point with $y = -1$.)

Are each of the following statements true or false? True means it always holds, with no other assumptions on the data set or model; false means that it need not hold.

- (a) Replacing v with zero will reduce, or not increase, the false positive rate.
- (b) Replacing β with zero will reduce, or not increase, the false positive rate.
- (c) Halving v (i.e., replacing v with $v/2$) will reduce, or at least not increase, the false positive rate.

- (d) Halving β (*i.e.*, replacing β with $(1/2)\beta$) will reduce, or at least not increase, the false positive rate.

A16.5 *Smallest fund holdings consistent with portfolio return history.* A fund consists of a portfolio with n assets, characterized by the weight n -vector w , which satisfies $\mathbf{1}^T w = 1$. (Negative entries correspond to short positions.) The fund manager publishes the return of the portfolio on T days, given by $r = R w$, where R is the $T \times n$ matrix of asset returns, but does not divulge her holdings w . (We do, however, know R .) We consider the case where T is smaller than n , *i.e.*, we are given fewer portfolio returns than there are assets. We will guess w by finding $\hat{w}$, the smallest weight vector (measured by norm squared) that is consistent with the reported portfolio returns (and satisfies $\mathbf{1}^T w = 1$).

Give a formula for $\hat{w}$ (in terms of R , r , and the dimensions n and T). Your formula can involve the usual matrix and vector operations, including matrix addition, transpose, multiplication, inverse, pseudo-inverse, and so on. (If you use inverse or pseudo-inverse, you can assume that the appropriate condition, such as linear independence of columns or rows, holds.)

A16.7 *De-biasing via linearly constrained least squares.* Prediction models are fit to historical data. When the historical data contains biases, these biases are captured in the model. If the model is then used to make future decisions, it can help to perpetuate the biases. As a specific example consider the problem of predicting interviewer ratings from attributes of a candidate that are culled from their resume. The model is fit using historical data, where due to conscious and unconscious bias, interviewers have given lower ratings to women on average. If the prediction model captures these biases, and is then used to screen applicants (which is a typical use), it will help perpetuate biases.

There is a whole field devoted to various aspects of this problem, such as detecting bias in data and models, and fitting data in ways that reduce the problem of perpetuating bias. In this exercise we examine a simple method based on ideas from the course.

We consider the specific case mentioned above. The historical (training) data consists of pairs $(x^{(i)}, y^{(i)})$, $i = 1, \dots, N$, where $x^{(i)}$ is an n -vector of features culled from a candidate's resume, and $y^{(i)}$ is the numerical rating given by the interviewer. We assume that the first N^f of the candidates are female and the remaining N^m candidates are male. (We have $N^f + N^m = N$, and we assume N^f and N^m are both positive.) We will fit a regression model to this data, *i.e.*, $\hat{y} = x^T \beta + v$, where the n -vector β and number v are the model parameters. We choose β and v by minimizing the sum square prediction error on the training data,

$$\sum_{i=1}^N (\hat{y}^{(i)} - y^{(i)})^2,$$

where $\hat{y}^{(i)} = (x^{(i)})^T \beta + v$. We also impose an *anti-bias constraint*, which is that the average predicted score for the female and male candidates on the training set are equal, *i.e.*,

$$\frac{1}{N^f} \sum_{i=1}^{N^f} \hat{y}^{(i)} = \frac{1}{N^m} \sum_{i=N^f+1}^N \hat{y}^{(i)}.$$

Explain how to formulate this as a linearly constrained least squares problem with variable (β, v) of the form

$$\begin{aligned} &\text{minimize} && \|A(\beta, v) - b\|^2 \\ &\text{subject to} && C(\beta, v) = d. \end{aligned}$$

Give A , b , C , and d , making sure to specify their dimensions and entries. You can use X , the $n \times N$ matrix of features, with columns $x^{(i)}$, and the N -vector y^d , with entries $y^{(i)}$. You are welcome to slice X and y into the female and male portions, *e.g.*, $X_f = X_{1:n, 1:N^f}$ or $y_m = y_{N^f+1:N}^d$.