Homework 0 B

Fall 2020, CSE 546: Machine Learning
John Franklin Crenshaw
October 5, 2020

Probability and Statistics

B.1 Let $X_1, \dots, X_n$ be n independent and identically distributed random variables drawn uniformly at random from $[0, 1]$. If $Y = \max\{X_1, \dots, X_n\}$ then find $\mathbb{E}[Y]$.

Let $F_X(x)$ be the CDF and $f_X(x)$ be the PDF of the uniform distribution from which X is drawn. As the distribution is uniform on 0 to 1, $f_X(x) = 1$ in this domain and zero elsewhere. Therefore,

$$F_X(x) = \int_{-\infty}^{x} f_X(x')dx' = \int_0^x dx' = x.$$

Now the probability that the max of $\{X_1, \dots, X_n\}$ is less than y is

$$F_Y(y) = P(\max\{X_1, \dots, X_n\} < y) = \prod_{i=1}^{n} P(X_i < y) = \prod_{i=1}^{n} F_X(y) = \prod_{i=1}^{n} y = y^n.$$

Taking the derivative of this CDF, we get the PDF:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = ny^{n-1}.$$

Finally, we can calculate $\mathbb{E}[Y]$:

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y)dy = \int_0^1 y \cdot ny^{n-1} = \frac{n}{n+1}$$

B.2 Let X be a random variable with $\mathbb{E}[X] = \mu$ and $\mathbb{E}[(X - \mu)^2] = \sigma^2$. For any $x \geq 0$, use Markov's inequality to show that $\mathbb{P}(X \geq \mu + \sigma x) \leq 1/x^2$.

Using Markov's Inequality, $\mathbb{P}(X \geq a) = \frac{\mathbb{E}[X]}{a}$, we have

$$\mathbb{P}(X \geq \mu + \sigma x) = \mathbb{P}[(X - \mu)^2 \geq \sigma^2 x^2] \leq \frac{\mathbb{E}[(X - \mu)^2]}{\sigma^2 x^2} = \frac{\sigma^2}{\sigma^2 x^2} = \frac{1}{x^2}.$$

Thus, $\mathbb{P}(X \geq \mu + \sigma x) \leq 1/x^2$.

Linear Algebra and Vector Calculus

B.3 The *trace* of a matrix is the sum of the diagonal entries; $\mathrm{Tr}(A) = \sum_i A_{ii}$. If $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times n}$, show that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.

$$\mathrm{Tr}(AB) = \sum_i [AB]_{ii} = \sum_i \left(\sum_j A_{ij} B_{ji} \right) = \sum_j \left(\sum_i B_{ji} A_{ij} \right) = \sum_j [AB]_{jj} = \mathrm{Tr}(BA)$$

B.4 Let $v_1, \dots, v_n$ be a set of non-zero vectors in $\mathbb{R}^d$. Let $V = [v_1, \dots, v_n]$ be the vectors concatenated.

a. What is the minimum and maximum rank of $\sum_{i=1}^{n} v_i v_i^T$?

Let $M = \sum_{i=1}^n v_i v_i^T$. As the vectors v_i are in $\mathbb{R}^d$, at most d of them can be linearly independent. Of course, if $n < d$, then at most n vectors can be linearly independent. If you imagine the case where $v_1, \ldots, v_n$ are drawn from an orthonormal basis of $\mathbb{R}^d$, then you can see that $\text{rank}(M) \leq \min(n, d)$. We can also imagine the case where all of the vectors are the same. Then $\text{rank}(M) = 1$. The rank cannot be zero, as the vectors are non-zero. Therefore, $1 \leq \text{rank}(M) \leq \min(n, d)$.

b. What is the minimum and maximum rank of V?

The argument from part a works here as well, just replacing N for V. Thus $1 \leq \text{rank}(V) \leq \min(n, d)$.

c. Let $A \in \mathbb{R}^{D \times d}$ for $D > d$. What is the minimum and maximum rank of $\sum_{i=1}^n (Av_i)(Av_i)^T$?

The resultant matrix is $AMA^T \in \mathbb{R}^{D \times D}$, where M is defined above. Despite being a matrix with D columns and rows, it cannot have greater rank than M. This can be seen via the same arguments given in part a, because acting on n vectors in $\mathbb{R}^d$ with the same linear transformation cannot result in more linearly independent vectors than you started with. However, multiplying by A can *reduce* the rank, as the image of A may be lower dimensional than the set of v_i. So $1 \leq \text{rank}(AMA^T) \leq \min(n, d, \text{rank}(A))$.

d. What is the minimum and maximum rank of AV? What if V is rank d?

Again, we can use the same argument as we did in part c. Thus $1 \leq \text{rank}(AV) \leq \min(n, d, \text{rank}(A))$.