STATS 202A – FINAL PROJECT TASK 1

NAME: ANOOSHA SAGAR

Student ID: 605028604

PROOF READING "A NOTE ON MULTIVARIATE CALCULUS AND MULTIVARIATE STATISTICS"

SECTION 1.6

1.6 Noun or verb, change of viewpoint

A matrix collects a bunch of numbers, and can be viewed as a noun. But most of the time, we multiple a matrix to a vector to transform it to another vector, hence a matrix is a verb. The meaning of the verb sometimes means a change of viewpoint. Thus a mathematical expression in terms of such a matrix can often be read in terms of English. That is, matrices give us a mathematical language that encodes rich meanings.

SPELLING MISTAKE: *IT SHOULD BE "**MULTIPLY**" INSTEAD OF "**MULTIPLE**"*

SECTION 2.1

2.1 First derivative

Suppose $Y = (y_i)_{m\times 1}$, and $X = (x_j)_{n\times 1}$. Suppose $Y = h(X)$. We can define

$$\frac{\partial Y}{\partial X^\top} = \left(\frac{\partial y_i}{\partial x_j}\right)_{m\times n}.$$

Here is the key. The above definition is not even necessary, because it follows directly from matrix multiplication. Specifically, we can treat $\partial Y = (\partial y_i, i = 1, ..., m)^\top$ as a column vector, and $1/\partial X = (1/\partial x_j, j = 1, ..., m)^\top$ as another column vector. Now we have two vectors of operations, instead of numbers. The product of the elements of the two vectors is understood as composition of the two operators, i.e., $\partial y_i(1/\partial x_j) = \partial y_i/\partial x_j$. Then $\partial Y/\partial X^\top$ is a squared matrix according to the matrix multiplication rule.

Should be ***j = 1,…,n***

SECTION 2.2

2.2 Second derivative

By the same reasoning, if Y is a scaler, then the gradient $h'(X) = \partial Y/\partial X$ is a $n \times 1$ column vector, and $\partial Y/\partial X^\top$ is a $1 \times n$ row vector. For scaler Y, we can define the Hessian or second derivative

$$h''(X) = \frac{\partial^2 Y}{\partial X \partial X^\top} = \left(\frac{\partial^2 Y}{\partial x_i \partial x_j}\right)_{n\times n}.$$

SPELLING MISTAKE: *IT SHOULD BE "**SCALAR**" INSTEAD OF "**SCALER**"*

SECTION 2.3

2.3 Examples

If $Y = AX$, then $y_i = \sum_k a_{ik}x_k$. Thus $\partial y_i/\partial x_j = a_{ij}$. So $\partial Y/\partial X^\top = A$.
If $Y = X^\top SX$, where S is symmetric, then $\partial Y/\partial X = 2SX$, and $\partial^2 Y/\partial X \partial X^\top = 2S$.
If $S = I$, $Y = \|X\|^2$, $\partial Y/\partial X = 2X$.
The above results generalize the scalar results with almost no change in notation.

SHOULD BE X^T IN THE DENOMINATOR TERM OF THE HIGHLIGHTED TEXT

SECTION 2.10

2.10 Orthogonal matrix: viewpoint

An orthogonal matrix $Q = (q_1, ..., q_n)$ is such as $\langle q_i, q_j \rangle = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. That is $(q_1, ..., q_n)$ form an orthonormal basis.

For a vector v, we can view it in Q, so that $v = q_1u_1 + ... + q_nu_n = (q_1, ..., q_n)u = Qu$. In the last step of the calculation, we may treat each q_i as a number in our mental calculation. Each $u_i = \langle v, q_i \rangle = q_i^\top u$. Thus $u = Q^\top u$, where again we can treat each $q_i^\top$ as a number in our mental calculation. Thus, v becomes $u = (u_1, ..., u_n)^\top$ from the point of view of Q. $u = Q^\top v$ is analysis, i.e., we decompose v into pieces along $(q_1, ..., q_n)$. $v = Qu$ is synthesis, i.e., we put the pieces together to get back v. Clearly $Q^\top = Q^{-1}$, i.e., $QQ^\top = Q^\top Q = I$.

The highlighted text should be:

$u_i = \langle u, q_i \rangle$

$v = Q^\top u$

SECTION 2.17

2.17 Least squares

Let the data frame be $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ is $n \times p$ and Y is $n \times 1$. The model is $\mathbf{Y} = \mathbf{X}\beta_{\text{true}} + \epsilon$, where β is $p \times 1$, and ϵ is $n \times 1$.

Let $R(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|^2$ be the least squares loss function, then

$$R'(\beta) = -2\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\beta)$$

and

$$R''(\beta) = 2\mathbf{X}^\top\mathbf{X}.$$

We can derive these by the chain rule. Let $e = \mathbf{Y} - \mathbf{X}\beta$. Then

$$\frac{\partial R}{\partial \beta^\top} = \frac{\partial R}{\partial e^\top}\frac{\partial e}{\partial \beta^\top} = -2e^\top\mathbf{X}.$$

$R'(\beta) = \partial R/\partial\beta$, which is obtained by transposing $-2e^\top\mathbf{X}$.

$$R''(\beta) = \frac{\partial^2 R}{\partial\beta\partial\beta^\top} = \partial(-2\mathbf{X}^\top e)/\partial\beta^\top = -2\mathbf{X}^\top\mathbf{X} < 0.$$

HIGHLIGHTED TEXT SHOULD BE $-2X^TX > 0$

SECTION 3.2

3.2 Variance of a random vector

Let X be a random vector. Let $\mu_X = \mathrm{E}(X)$. We define

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mu_X)(X - \mu_X)^\top].$$

Then the (i, j)-th element of $\mathrm{Var}(X)$ is $\mathrm{Cov}(x_i, x_j)$. The diagonal elements are $\mathrm{Var}(x_i)$.

Let A be a constant matrix of appropriate dimension, then

$$\mathrm{Var}(AX) = A\mathrm{Var}(X)A^\top.$$

This is because

$$\begin{aligned}
\mathrm{Var}(AX) &= \mathrm{E}[(AX - \mathrm{E}(AX))(AX - \mathrm{E}(AX))^\top] \\
&= \mathrm{E}[(AX - A\mu_X)(AX - A\mu_X)^\top] \\
&= \mathrm{E}[A(X - \mu_X)(X - \mu_X)^\top A^\top] \\
&= A\mathrm{E}[(X - \mu_X)(X - \mu_X)^\top]A^\top \\
&= A\mathrm{Var}(X)A^\top.
\end{aligned}$$

THE HIGHLIGHTED TEXT CAN BE OMITTED

SECTION 3.5

3.5 Principal component analysis

Assuming $\mathrm{E}(X) = 0$ (otherwise we can let $X \leftarrow X - \mathrm{E}(X)$), and $\mathrm{Var}(X) = \Sigma = Q\Lambda Q^\top$. Then viewed from Q, $\mathrm{E}(Z) = 0$ and $\mathrm{Var}(Z) = \Lambda$.

Assume $\lambda_1 > \lambda_2 > ... > \lambda_n > 0$. If $\lambda_i = \mathrm{Var}(z_i)$ is very small for $i > m$, then $z_i \approx 0$ for $i > m$ (recall $\mathrm{E}(z_i) = 0$). We can represent

$$X \approx \sum_{i=1}^{m} q_i z_i,$$

thus reducing the dimensionality of X from n to m. The $(q_i, i = 1, ..., m)$ are called principal components.

For instance, if X is a face image, then $(q_i, i = 1, ..., m)$ are the eigen faces, which may correspond to different features of a face (e.g., eyes, nose, mouth etc.), and $(z_i, i = 1, ..., m)$ is a low dimensional representation of X.

ALTHOUGH X HAS BEEN DEFINED IN THE PREVIOUS SECTION, IT SHOULD HAVE BEEN REDEFINED HERE TO MAINTAIN CONTINUITY

SECTION 3.7

3.7 Multivariate normal

We start from $Z = (z_1, ..., z_n)^\top$, where $z_i \sim \mathrm{N}(0, 1)$ independently. Then $\mathrm{E}(Z) = 0$, and $\mathrm{Var}(Z) = I$. We denote $Z \sim \mathrm{N}(0, I)$. The density of Z is

$$f_Z(Z) = \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_i z_i^2\right] = \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2} Z^\top Z\right].$$

Let $X = \mu + \Sigma^{1/2} Z$, then $Z = \Sigma^{-1/2}(X - \mu)$, which is a matrix version of standardization. Then

$$\begin{aligned} f_Y(Y) &= \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}(X-\mu)^\top \Sigma^{-1}(X-\mu)\right] / |\Sigma^{1/2}| \\ &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(X-\mu)^\top \Sigma^{-1}(X-\mu)\right]. \end{aligned}$$

THE HIGHLIGHTED TEXT SHOULD BE REPLACED WITH $F_Y(X)$