Statistical Learning 2024 HW ANSWERS

Brad Baxter

20240321

1. (i). The SVD is the factorization $A = USV^T$, where $U \in O(m)$, $V \in O(n)$ and $S \in \mathbb{R}^{m\times n}$ is a diagonal matrix whose diagonal elements satisfy

$$s_1 \geq s_2 \geq \cdots \geq s_n.$$

The diagonal elements of S are called the singular values of A.

3 pts

(ii). Given any pair of matrices $A, B \in \mathbb{R}^{m\times n}$, their Frobenius inner product is given by

$$\langle A, B\rangle_F = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}B_{ij}.$$

The Frobenius norm is defined by

$$\|A\|_F = \sqrt{\langle A, A\rangle_F}.$$

4 pts

(iii). If $A = (\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_n)$, then

$$\|QA\|_F^2 = \sum_{k=1}^{n} \|Q\mathbf{a}_k\|_2^2 = \sum_{k=1}^{n} \|\mathbf{a}_k\|_2^2 = \|A\|_F^2,$$

because an orthogonal matrix leaves the Euclidean norm of a vector unchanged. Now

$$\|AR\|_F = \|\left(AR\right)^T\|_F = \|R^TA^T\|_F = \|A^T\|_F = \|A\|_F,$$

because $R \in O(n)$ if and only if $R^T \in O(n)$, and the Frobenius norm is invariant under the transpose operation, which is obvious from its definition.

4 pts

(iv). We have

$$\|A - Q\|_F^2 = \|USV^T - Q\|_F^2 = \|U^T\left(USV^T - Q\right)V\|_F^2 = \|S - U^TQV\|_F^2,$$

since the Frobenius norm is invariant under pre- and post-multiplication by orthogonal matrices. Thus

$$\|A - Q\|_F^2 = \|S - W\|^2 = \langle S - W, S - W\rangle_F = \|S\|_F^2 - 2\langle S, W\rangle_F + \|W\|_F^2.$$

Now every column of an orthogonal matrix is a unit vector, which implies $\|W\|_F^2 = n$. Further, since S is a diagonal matrix, $\langle S, W\rangle_F = s_1W_{11} + \cdots + s_nW_{nn}$. Therefore

$$\|A - Q\|_F^2 = \|S\|_F^2 - 2\sum_{k=1}^{n} s_kW_{kk} + n = \sum_{k=1}^{n} s_k^2 - 2s_kW_{kk} + 1.$$

4 pts

(v). We have

$$\|A - Q\|_F^2 = \sum_{k=1}^{n} s_k^2 + 1 - 2\sum_{k=1}^{n} s_k W_{kk}.$$

Thus minimizing $\|A-Q\|_F$ is equivalent to maximizing $\sum_{k=1}^{n} s_k W_{kk}$, for $W \in O(n)$. Now every column of an orthogonal matrix is a unit vector, so its diagonal elements satisfy $-1 \leq W_{kk} \leq 1$. Hence

$$\sum_{k=1}^{n} s_k W_{kk} \leq \sum_{k=1}^{n} s_k,$$

with equality if $U^T Q V = W = I$, or $Q = UV^T$.

The Procrustes problems arises in many areas, but one possible application is in missile guidance systems, where A is a perturbed orthogonal matrix, generated by hardware, which specifies the orientation of the missile.

5 pts

2. (i). Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be points in $\mathbb{R}^d$. The k-means algorithm is a simple method for iteratively updating a set of k *cluster centres* $\mathbf{m}_1, \ldots, \mathbf{m}_k$. At the start of the algorithm, these points can be any vectors.

Now the k cluster centres partition $\mathbb{R}^d$ into k clusters: we let the ith cluster C_i be those points in $\mathbb{R}^d$ for which $\mathbf{m}_i$ is the closest cluster centre, that is

$$C_i = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{m}_i\| = \min_{1 \le \ell \le k} \|\mathbf{x} - \mathbf{m}_\ell\|\}, \qquad 1 \le i \le n,$$

and students are not expected to deal with ambiguous cases for which some points lie in more than one cluster. We then replace each cluster centre $\mathbf{m}_i$ by the centroid of the subset of points in $\mathbf{x}_1, \ldots, \mathbf{x}_n$ which are contained in the ith-cluster (the centroid of a finite set of points $\mathbf{v}_1, \ldots, \mathbf{v}_j$ is simply the sample average $(\mathbf{v}_1 + \cdots + \mathbf{v}_j)/j$). The new cluster centres then define corresponding new centres, and we then repeat the procedure until the cluster centres converge.

8 pts

(ii). We can summarize the links between websites by a single matrix containing 0s and 1s. Specifically, if there are N websites, then we let $W_{ij} = 1$ if site i links to site j and $i \neq j$, but otherwise set $W_{ij} = 0$. A

Page and Brin decided to rank these N websites by simulating user behaviour with a Markov model based on the connectivity matrix W. Specifically, we imagine vast numbers of users surfing the web in discrete time. At the kth step, the vector $\pi^{(k)}$ denotes the probability distribution for our users, that is, $\pi_i^{(k)}$ is the probability that a user is surfing site i at time k. We then let our users surf to new sites according to the transition matrix $P \in \mathbb{R}^{N \times N}$, where

$$P_{ij} = \frac{W_{ij}}{\sum_{k=1}^N W_{ik}}, \qquad 1 \le i, j \le N. \tag{1}$$

Further, we shall assume that $\sum_{k=1}^n W_{ik} \neq 0$, for all i, to avoid a zero denominator in the definition of P (we are assuming that there are no *dangling pages*, to use Google's jargon).

Thus the new probability vector is given by

$$\pi^{(k+1)} = P^T \pi^{(k)} \tag{2}$$

and, over time, we hope to obtain an *invariant measure* (or stationary probability vector) π. Unfortunately this Markov chain turns out to be inadequate, because most sites tend to fall into isolated clusters and it inherits this stagnation. One way to avoid this is a *teleporting random walk*: we choose a parameter $c \in (0, 1)$ and *either* use P with probability c, *or* move to one of the N websites with equal probability. Thus our new transition matrix is

$$M = cP + (1 - c)\frac{\mathbf{e}\mathbf{e}^T}{N}, \tag{3}$$

where

$$\mathbf{e} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{4}$$

The new invariant measure vector π now satisfies $M^T\pi = \pi$.

Page and Brin decided to define the *rank vector* $\mathbf{r} = N\pi$. Thus the last equation becomes

$$\left(I - cP^T\right)\mathbf{r} = (1-c)\mathbf{e}. \tag{5}$$

This linear system contains N linear equations in N unknowns, but $N \approx 10^9$. Unfortunately, direct elimination requires $T(N) = CN^3$ seconds, where $T(10^3) \approx 1$ on basic modern computer. Hence elimination is completely unsuitable. Fortunately, a simple iterative algorithm called *Jacobi's method* is available. Specifically, given any $n \times n$ matrix A, Jacobi's method attempts to solve $A\mathbf{x} = \mathbf{y}$ as follows. We first choose any initial vector $\mathbf{x}^{(0)}$. Then, given $\mathbf{x}^{(k-1)}$, we define $\mathbf{x}^{(k)}$ by the equation

$$x_i^{(k)} = \frac{y_i}{A_{ii}} - \sum_{j=1, j\neq i}^{n} \left(\frac{A_{ij}}{A_{ii}}\right) x_j^{(k)}, \qquad 1 \le i \le n. \tag{6}$$

Hence

$$\mathbf{r}^{(k)} = cP^T\mathbf{r}^{(k-1)} + (1-c)\mathbf{e}. \tag{7}$$

12 pts

3. (i). We have, recalling that $S_{pq} = s_p \delta_{pq}$,

$$\begin{aligned} A_{ij} &= \sum_{p=1}^{m} U_{ip} (SV^T)_{pj} \\ &= \sum_{p=1}^{m} \sum_{q=1}^{n} U_{ip} S_{pq} V_{jq} \\ &= \sum_{p=1}^{n} s_p U_{ip} V_{jp} \\ &= \sum_{p=1}^{n} s_p \mathbf{u}_p(i) \mathbf{v}_p(j) \\ &= \left(\sum_{p=1}^{n} s_p \mathbf{u}_p \mathbf{v}_p^T \right)_{ij}, \end{aligned}$$

as required. **6 pts**

(ii). We have

$$A_r \mathbf{v}_\ell = \sum_{k=1}^{r} s_k \mathbf{u}_k \mathbf{v}_k^T \mathbf{v}_\ell = 0,$$

if $\ell > r$. **3 pts**

(iii). We have

$$A_r \mathbf{x} = \sum_{k=1}^{r} s_k (\mathbf{v}_k^T \mathbf{x}) \mathbf{u}_k.$$

3 pts

(iv). The orthogonal invariance of the Frobenius norm implies

$$\|A - A_r\|_F^2 = \|S - S_r\|_F^2 = s_{r+1}^2 + \cdots + s_n^2,$$

where $S_r = \text{diag}\,\{s_1, \ldots, s_r, 0, \ldots, 0\}$.

3 pts

(v). We have $\|(A - A_r)\mathbf{x}\|_2 = \|(S - S_r)\mathbf{y}\|_2$, where $\mathbf{y} = V^T \mathbf{x}$ and $\|\mathbf{y}\|_2 = \|\mathbf{x}\|_2$. Now

$$\|(S - S_r)\mathbf{y}\|_2^2 = s_{r+1}^2 y_{r+1}^2 + \cdots + s_n^2 y_n^2 \le s_{r+1}^2 \|\mathbf{y}\|^2,$$

because $s_1 \ge \cdots \ge s_n$. Hence $\|(A - A_r)\mathbf{x}\|_2 \le s_{r+1} \|\mathbf{x}\|_2$, as required.

5 pts