Data Mining Coursework

Brad Baxter

20070228

Due Tuesday, April 24th, to the General Office (**not** to me). Attempt all questions.

1. Let A be any real $m \times n$ matrix, where $m \geq n$. Throughout this question $O(k)$ will denote the set of real $k \times k$ orthogonal matrices. You may assume standard properties of orthogonal matrices if clearly stated.

   (i). Define the singular value decomposition.

   3 pts

   (ii). Define the Frobenius norm and inner product.

   4 pts

   (iii). Prove that $\|QA\|_F = \|AR\|_F = \|A\|_F$, for any $A \in \mathbb{R}^{m \times n}$, $Q \in O(m)$ amd $R \in O(n)$.

   4 pts

   (iv). Given any matrix $A \in \mathbb{R}^{n \times n}$, with singular value decomposition $A = USV^T$, prove that
   $$\|A - Q\|_F^2 = \|S - W\|_F^2,$$
   for any $Q \in O(n)$, where $W = U^T Q V$. Hence show that
   $$\|A - Q\|_F^2 = \sum_{k=1}^{n} \left(s_k^2 - 2 s_k W_{kk} + 1 \right),$$
   where $s_1, \ldots, s_n$ are the singular values of A.

   4 pts

   (v). Hence, or otherwise, prove that the non-linear least squares Procrustes problem
   $$\min_{Q \in O(n)} \|A - Q\|_F,$$
   where $A \in \mathbb{R}^{n \times n}$, is solved by setting $Q = UV^T$. Briefly describe one application of this Procrustes problem.

   5 pts

2. (i). Describe the k-means clustering algorithm.

8 pts

(ii). Describe the PageRank algorithm. You should describe a suitable iterative method for obtaining the PageRank stationary distribution, but need not prove convergence.

12 pts

3. The Gamma function is defined by the integral relation

$$\Gamma(\alpha) = \int_0^\infty e^{-s} s^{-1+\alpha}\, ds, \qquad \text{for } \alpha > 0.$$

(i). Use the substitution $s = r^2 t$ to show that

$$r^{-2\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^\infty e^{-r^2 t} t^{-1+\alpha}\, dt,$$

for any positive r.

5 pts

(ii). Hence, or otherwise, prove that

$$\frac{1}{(\|\mathbf{x}\|_2^2 + c^2)^\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^\infty e^{-\|\mathbf{x}\|_2^2 t} e^{-c^2 t} t^{-1+\alpha}\, dt,$$

for any $\mathbf{x} \in \mathbb{R}^d$ and positive constant c.

5 pts

(iii). Prove that the Radial Basis Function

$$s(\mathbf{x}) = \sum_{j=1}^n \frac{a_j}{(\|\mathbf{x} - \mathbf{x}_j\|_2^2 + c^2)^\alpha}, \qquad \mathbf{x} \in \mathbb{R}^d,$$

can be used to interpolate arbitrary observations $s(\mathbf{x}_i) = f_i$, for $1 \le i \le n$, when the points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ are distinct.

10 pts

[You may assume that

$$\sum_{j=1}^n \sum_{k=1}^n a_j a_k \exp(-\lambda \|\mathbf{x}_j - \mathbf{x}_k\|_2^2) \ge 0,$$

for any real numbers $a_1, \ldots, a_n$ and any vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, where λ can be any positive constant. You may also assume that this inequality is strict when the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are distinct and the coefficients $a_1, \ldots, a_n$ are not identically zero.]

4. Let $A \in \mathbb{R}^{m\times n}$, where $m \geq n$, have singular value decomposition $A = USV^T$, where $U = (\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_m)$, $V = (\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_n)$ and the singular values $s_1, \ldots, s_n$ are all positive.

(i). Prove that

$$A = \sum_{k=1}^{n} s_k \mathbf{u}_k \mathbf{v}_k^T.$$

6 pts

(ii). For $1 \leq r \leq n$, define

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

Prove that $A_r\mathbf{w} = 0$ if $\mathbf{w} \in \text{span}\,\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$.

3 pts

(iii). Prove that $A_r\mathbf{x} \in \text{span}\,\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$, for any $\mathbf{x} \in \mathbb{R}^n$.

3 pts

(iv). Calculate $\|A - A_r\|_F^2$, for $1 \leq r \leq n$.

3 pts

(v). Prove that $\|(A - A_r)\mathbf{x}\|_2 \leq s_{r+1}\|\mathbf{x}\|_2$, for $1 \leq r \leq n$.

5 pts

5. Let $f : [0, \infty) \to \mathbb{R}$ be a function with the property that the quadratic form

$$Q := \sum_{j=1}^{n} \sum_{k=1}^{n} a_j a_k f(\|\mathbf{x}_j - \mathbf{x}_k\|_2^2)$$

is always non-negative, for *any* dimension d, for any number n of points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, and for any real numbers $a_1, \ldots, a_n \in \mathbb{R}$.

(i). Prove that $f(0) \geq 0$.

4 pts

(ii). Let λ be any nonzero real number, let $a_j = 1$, for $1 \leq j \leq n$, and let $\mathbf{x}_j = \lambda \mathbf{e}_j$, for $j = 1, \ldots, n$, where $\mathbf{e}_j$ is the jth coordinate vectors. Prove that

$$Q = nf(0) + n(n-1)f(2\lambda^2).$$

10 pts

(iii). Hence prove that $f(t) \geq 0$, for every $t > 0$.

6 pts

6. One way to use radial basis functions in data mining is as follows. We are given sequences of points $\mathbf{b}_1, \ldots, \mathbf{b}_m$ and $\mathbf{c}_1, \ldots, \mathbf{c}_n$ lying in $\mathbb{R}^d$ and, given function values $f_1, \ldots, f_n$, we seek real coefficients $a_1, \ldots, a_m$ minimizing the sum of squares

$$\sum_{\ell=1}^{n} \left(f_\ell - s(\mathbf{c}_\ell)\right)^2,$$

where

$$s(\mathbf{x}) = \sum_{k=1}^{n} a_k \phi(\mathbf{x} - \mathbf{b}_k),$$

for some radially symmetric function $\phi : \mathbb{R}^d \to \mathbb{R}$. One solution to this problem, as for any least squares problem, is to solve the *normal equations*, which are given by

$$A^T A\mathbf{a} = A^T \mathbf{f},$$

where $\mathbf{a} = (a_1, \ldots, a_m)^T$, $\mathbf{f} = (f_1, \ldots, f_n)^T$ and

$$A_{\ell k} = \phi(\mathbf{c}_\ell - \mathbf{b}_k), \qquad 1 \le k \le m, \quad 1 \le \ell \le n.$$

(i). Show that

$$(A^T A)_{jk} = \sum_{\ell=1}^{n} \phi(\mathbf{c}_\ell - \mathbf{b}_j)\phi(\mathbf{c}_\ell - \mathbf{b}_k), \qquad 1 \le j, k \le m.$$

6 pts

(ii). Hence derive

$$\mathbf{v}^T A^T A\mathbf{v} = \sum_{\ell=1}^{n} \left(\sum_{k=1}^{m} v_k \phi(\mathbf{c}_\ell - \mathbf{b}_k) \right)^2.$$

6 pts

(iii). Now suppose that $\phi(\mathbf{x}) = e^{-\lambda\|\mathbf{x}\|^2}$, for $\mathbf{x} \in \mathbb{R}^d$, λ being a positive constant. Further, suppose that $\mathbf{b}_k^T \mathbf{c}_\ell = 0$, for all k and ℓ. Prove that there is a nonzero vector $\mathbf{v}$ for which $\mathbf{v}^T A^T A\mathbf{v} = 0$. Thus the matrix for the normal equations can be singular in some special cases.

8 pts