Data Mining Examination Questions

Brad Baxter

January 13, 2006

1. Let A be any real $m \times n$ matrix, where $m \geq n$.

   (i). Define the **singular value decomposition** $A = USV^T$.

   4 pts

   (ii). Prove that

   $$\|A\mathbf{x} - \mathbf{y}\|^2 = \|S\mathbf{a} - \mathbf{b}\|^2,$$

   where $\mathbf{a} = V^T\mathbf{x}$ and $\mathbf{b} = U^T\mathbf{y}$.

   4 pts

   (iii). If every singular value $s_1, \ldots, s_n$ of A is strictly positive, prove that the **least squares solution** $\mathbf{x}^*$ minimizing $\|A\mathbf{x} - \mathbf{y}\|^2$ is given by

   $$\mathbf{x}^* = VTU^T\mathbf{y},$$

   where $T \in \mathbb{R}^{m \times n}$ matrix whose only nonzero elements are given by

   $$T_{jj} = \frac{1}{s_j}, \qquad \text{for } 1 \leq j \leq n.$$

   4 pts

   (iv). Use the singular value decomposition to prove that the least squares solution vector is also given by

   $$\mathbf{x}^* = (A^TA)^{-1}A^T\mathbf{y}.$$

   8 pts

2. Briefly describe the k-means clustering algorithm.

8pts

Suppose we apply the k-means clustering algorithm when $k = 2$ and there are 4 points $(\pm R, \pm 1)$ in $\mathbb{R}^2$, where $R \gg 1$. Find the limiting clusters for the following initial centroid positions.

(i). $\mathbf{m}_1 = (-a, 0)$, $\mathbf{m}_2 = (a, 0)$, where $a > 0$. **4 pts**

(ii). $\mathbf{m}_1 = (0, -a)$, $\mathbf{m}_2 = (0, a)$, where $a > 0$. **4 pts**

(iii). $\mathbf{m}_1 = \mathbf{u}$, $\mathbf{m}_2 = -\mathbf{u}$, where $\mathbf{u}$ can be any unit vector. [You may ignore any cases when cluster membership is ambiguous.] **4 pts**

3. The theory of the Gamma function implies the equation

$$\Gamma(n+1/2) = \int_0^\infty e^{-t} t^{n-1/2}\, dt,$$

for any positive integer n.

(i). Use the change of variable $t = a^2 s$, where a is a positive constant, to show that

$$\frac{1}{a^{2n+1}} = \frac{1}{\Gamma(n+1/2)} \int_0^\infty e^{-a^2 s} s^{n-1/2}\, ds.$$

6 pts

(ii). Use the previous integral to derive

$$(r^2 + c^2)^{-(2n+1)/2} = \int_0^\infty e^{-r^2 s} w(s)\, ds, \qquad \text{for } r \geq 0,$$

where

$$w(s) = \frac{1}{\Gamma(n+1/2)} e^{-c^2 s} s^{n-1/2}$$

and c is a positive constant.

6 pts

(iii). Hence, or otherwise, prove that the radial basis function $\phi(r) = (r^2 + c^2)^{-(2n+1)/2}$, for $r \geq 0$, can be used for interpolation in dimension d. [You may use the fact that, if $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ are distinct vectors, then

$$\sum_{j=1}^{n} \sum_{k=1}^{n} v_j v_k e^{-s\|\mathbf{x}_j - \mathbf{x}_k\|^2} \geq 0$$

for $s > 0$, with equality if and only if $v_1 = \cdots = v_n = 0$.]

8 pts

4. Define the **Frobenius norm** $\|A\|_F$ of a matrix $A \in \mathbb{R}^{m\times n}$.

2 pts

Prove that $\|UAV\|_F = \|A\|_F$ when $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices.

6 pts

Briefly describe how the singular value decomposition can be used to solve the following nonlinear least-squares problems. [You do **not** need to prove that these solutions are valid.]

(i). Given any square matrix $A \in \mathbb{R}^{n\times n}$, find the closest orthogonal matrix Q_A, in the sense that

$$\|A - Q_A\|_F = \min_{Q\in O(n)} \|A - Q\|_F,$$

where $O(n)$ denotes the set of all real, $n \times n$ orthogonal matrices. [This is the square Procrustes problem.]

4 pts

(ii). Given matrices $A, B \in \mathbb{R}^{m\times n}$, $m \geq n$, find the orthogonal matrix $\hat{Q} \in O(n)$ minimizing

$$\|A - B\hat{Q}\|_F = \min_{Q\in O(n)} \|A - BQ\|_F.$$

[This is the rectangular Procrustes problem.] **4 pts**

(iii). Define the **rank** of a matrix $A \in \mathbb{R}^{m\times n}$, for $m \geq n$, in terms of the singular values of A. Find the closest matrix A_r of rank r to a general matrix A of rank n, in the sense that

$$\|A - A_r\|_F = \min_{\text{rank } B=r} \|A - B\|_F.$$

4 pts