DS-GA 1003: Homework 4
Kernels, Duals, and Trees

Due on Tuesday, March 22, 2016

Professor David Ronsenberg

See complete code at: *git@github.com:cryanzpj/1003.git*

Yuhao Zhao
Yz3085

2 Positive Semidefinite Matrices

– 2.1.

Let $A = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$, $A^TA = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and A is not symmetric.

– 2.2.

Since M is psd, we assume M is real and symmetric, thus by Spectral Theorem, we have

$$M = Q\Sigma Q^T$$

where Q is an orthogonal matrix ($Q^T = Q^{-1}$), Σ is diagonal.

$$\Sigma = Q^{-1}M(Q^T)^{-1} = Q^TMQ = \begin{pmatrix} q_1^TMq_1 & q_1^TMq_2 & \cdots & q_1^TMq_n \\ q_2^TMq_1 & q_2^TMq_2 & \cdots & q_2^TMq_n \\ \vdots & \vdots & \cdots & \vdots \\ q_d^TMq_1 & q_d^TMq_2 & \cdots & q_d^TMq_n \end{pmatrix}.$$

Since M is psd, $q_i^TMq_i \geq 0$, the diagonals of Σ are eigenvalues of M and are non-negative.

– 2.3.

i). If we have M $= BB^T$ for some B, for $\forall v \in \Re^n$

$$v^TMv = v^TBB^Tv = (B^Tv)^T(B^Tv) = ||B^Tv|| \geq 0$$

Therefore, M is psd
ii). If we know M is psd, by spectral theorem

$$M = Q\Sigma Q^T = Q\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}Q^T = Q\Sigma^{\frac{1}{2}}(Q\Sigma^{\frac{1}{2}})^T = BB^T$$

where $B = Q\Sigma^{\frac{1}{2}}$
$\Sigma^{\frac{1}{2}}$ is a diagonal matrix whose diagonal equals the sqrt root of $diag(\Sigma)$

This proves a symmetric matrix M can be expressed as $M = BB^T$ iff M is psd

3 Positive Definite Matrices

– 3.1.

M is pd, by Spectral Theorem,

$$M = Q\Sigma Q^T$$

$$\Sigma = Q^{-1}M(Q^T)^{-1} = Q^TMQ = \begin{pmatrix} q_1^TMq_1 & q_1^TMq_2 & \cdots & q_1^TMq_n \\ q_2^TMq_1 & q_2^TMq_2 & \cdots & q_2^TMq_n \\ \vdots & \vdots & \cdots & \vdots \\ q_d^TMq_1 & q_d^TMq_2 & \cdots & q_d^TMq_n \end{pmatrix}.$$

Since M is pd, $q_i^TMq_i > 0$, the diagonals of Σ are eigenvalues of M and are positive .

– **3.2.**

since M is positive definite, $M = Q\Sigma Q^T$

$$Q\Sigma Q^T M = Q\Sigma^{-1}Q^T Q\Sigma Q^T$$

Q is an orthogonal matrix, $Q^T Q = I$

$$RHS = Q\Sigma^{-1}\Sigma Q^T = QQ^T = I$$

Therefore, $Q\Sigma Q^T$ is the inverse of M

– **3.3.**

M is psd and symmetric, for $\forall v \in \Re^n, v \neq \vec{0}$, and $\lambda > 0$

$$v^T(M + \lambda I)v = v^T Mv + \lambda v^T v > 0$$

since $v^T Mv \geq 0, \lambda v^T v > 0$. Therefore, $v^T(M + \lambda I)v$ is positive definite.
To show, $M + \lambda I$ is symmetric, we know that $\forall i \neq j, (M + \lambda I)_{i,j} = M_{i,j} = M_{j,i} = (M + \lambda I)_{j,i}$. Thus $M + \lambda I$ is also symmetric.

let $v_1, ..., v_n, \lambda_1, ..., \lambda_n$ be the n eigenvalues and eigenvectors of M

$$(M + \lambda I)v_i = Mv_i + \lambda v_i = \lambda_i v_i + \lambda v_i = (\lambda_i + \lambda)v_i$$

Therefore, v_i is also a eigenvector of $M + \lambda I$ with corresponding eigenvalue equals to $(\lambda_i + \lambda)$.
$M + \lambda I = Q\Sigma Q^T, Q = \{v_1, ..., v_n\}, \Sigma_{i,i} = \lambda_i + \lambda$, Then we have

$$(M + \lambda I)^{-1} = (Q^T)^{-1}\Sigma^{-1}Q^{-1} = Q\Sigma^{-1}Q^T = \sum_{i=1}^{n} \frac{1}{\lambda_i + \lambda} v_i v_i^T$$

– **3.4.**

M is symmetric psd and N is symmetric pd, $\forall v \in \Re^n, v \neq \vec{0}$

$$v^T(M + N)v = v^T Mv + v^T Nv$$

we know $v^T Mv \geq 0, v^T Nv > 0$

$$v^T(M + N)v > 0$$

This shows $M + N$ is positive definite.
To show $M + N$ is symmetric, $\forall i \neq j, (M + N)_{i,j} = M_{i,j} + N_{i,j} = M_{j,i} + N_{j,i} = (M + N)_{j,i}$, Thus $M + N$ is also symmetric. From 3.2 we know that positive definite matrix has inverse. Therefore, M+N is invertible.

4 Kernel Matrices

$$K = XX^T = \begin{pmatrix} x_1^T x_1 & \cdots & x_1^T x_m \\ \vdots & \vdots & \vdots \\ x_m^T x_1 & \cdots & x_m^T x_m \end{pmatrix}$$

$d(x_i, x_j) = ||x_i - x_j|| = \sqrt{(x_i - x_j) \cdot (x_i - x_j)} = \sqrt{x_i \cdot x_i + x_j \cdot x_j - 2x_i \cdot x_j} = \sqrt{K_{i,i} + K_{j,j} - 2K_{i,j}}$
Therefore , knowing K is equivalent to knowing the set of pairwise distance of vectors in S.

5 Kernel Ridge Regression

– 5.1.

Since

$$J(w) = ||Xw - y|| + \lambda||w^2|| \tag{1}$$

$$\frac{\partial J}{\partial w} = 2X^T(Xw - y) + 2\lambda wI = 0 \tag{2}$$

we have

$$X^T Xw - X^T y + \lambda wI = (X^T X + \lambda I)w - X^T y = 0 \tag{3}$$

$$w^* = (X^T X + \lambda I)^{-1} X^T y \tag{4}$$

XX^T is positive semidefinite and $\lambda > 0$, by 3.3, $XX^T + \lambda I$ is positive definite, thus invertible.

– 5.2.

Since $X^T Xw + \lambda Iw = X^T y, w = \frac{1}{\lambda}(X^T y - X^T Xw) = X^T \frac{1}{\lambda}(y - Xw)$
Thus $w = X^T \alpha$, where $\alpha = \frac{1}{\lambda}(y - Xw)$

– 5.3.

Since $w = X^T \alpha = \sum_1^n \alpha_i x_i$, w is a linear combination of data vectors

– 5.4.

since $w = X^T \alpha$ and $X^T Xw + \lambda Iw = X^T y$

$$X^T XX^T \alpha + \lambda IX^T \alpha = X^T y \tag{5}$$

$$X^T(XX^T + \lambda I)\alpha = X^T y \tag{6}$$

Therefore $\alpha = (XX^T + \lambda I)^{-1} y$

– 5.5.

Since $w = X^T \alpha = X^T(XX^T + \lambda I)^{-1} y, XX^T = K$

$$Xw = XX^T(XX^T + \lambda I)^{-1} y \tag{7}$$

$$= K(K + \lambda I)^{-1} y \tag{8}$$

– 5.6.

For a new point $\tilde{x}$

$$\tilde{x}^T w^* = \tilde{x}^T X^T (K + \lambda I)^{-1} y \tag{9}$$

$$= \begin{pmatrix} \tilde{x}^T x_1 & \tilde{x}^T x_2 & \cdots & \tilde{x}^T x_n \end{pmatrix} (K + \lambda I)^{-1} y \tag{10}$$

$$= k_{\tilde{x}}^T (K + \lambda I)^{-1} y \tag{11}$$

6 Decision Trees

• 6.1 Building Trees by Hand.

– 6.1.1.

a) Split on size:
i) Size ≤ 1, $p_1 = \frac{2}{3}, N_1 = 3, Q_1 = \frac{4}{9}, p_2 = \frac{3}{8}, N_2 = 8, Q_2 = \frac{30}{64}, N_1Q_1 + N_2Q_2 = \frac{61}{12} \approx 5.08$
ii) Size ≤ 2, $p_1 = \frac{2}{5}, N_1 = 5, Q_1 = \frac{12}{25}, p_2 = \frac{3}{6}, N_2 = 6, Q_2 = \frac{18}{36}, N_1Q_1 + N_2Q_2 \approx 5.4$
iii) Size ≤ 3, $p_1 = \frac{2}{6}, N_1 = 6, Q_1 = \frac{16}{36}, p_2 = \frac{3}{5}, N_2 = 5, Q_2 = \frac{12}{25}, N_1Q_1 + N_2Q_2 \approx 5.06$
iv) Size ≤ 4, $p_1 = \frac{4}{9}, N_1 = 9, Q_1 = \frac{40}{81}, p_2 = \frac{1}{2}, N_2 = 2, Q_2 = \frac{1}{2}, N_1Q_1 + N_2Q_2 \approx 5.4$

b) split on spots:
v) spots = N, $p_1 = 0, N_1 = 4, Q_1 = 0, p_2 = \frac{5}{7}, N_2 = 7, Q_2 = \frac{20}{49}, N_1Q_1 + N_2Q_2 \approx 2.85$

c) split on color:
vi) color = white, $p_1 = \frac{2}{5}, N_1 = 5, Q_1 = \frac{12}{25}, p_2 = \frac{3}{6}, N_2 = 6, Q_2 = \frac{18}{36}, N_1Q_1 + N_2Q_2 \approx 5.4$
The minimal weighted impurity measure is obtained by splitting on the spots.

Original

Spots = N

Poisonous	Size	Spots	Color
N	5	N	White
N	2	N	Brown
N	4	N	White
N	1	N	Brown

Spots = Y

Poisonous	Size	Spots	Color
N	2	Y	White
N	3	Y	Brown
Y	5	Y	White
Y	4	Y	Brown
Y	4	Y	Brown
Y	1	Y	White
Y	1	Y	Brown

– 6.1.2.

Since the left node is already pure, we continue splitting on the right node.

a) Split on color:
i) color = white, $p_1 = \frac{2}{3}, N_1 = 3, Q_1 = \frac{4}{9}, p_2 = \frac{3}{4}, N_2 = 4, Q_2 = \frac{6}{16}, N_1Q_1 + N_2Q_2 \approx 2.83$

b) Split on Size:
ii) Size ≤ 1, $p_1 = 0, N_1 = 2, Q_1 = 0, p_2 = \frac{3}{5}, N_2 = 5, Q_2 = \frac{12}{25}, N_1Q_1 + N_2Q_2 \approx 2.4$
iii) Size ≤ 2, $p_1 = \frac{2}{3}, N_1 = 3, Q_1 = \frac{4}{9}, p_2 = \frac{3}{4}, N_2 = 4, Q_2 = \frac{6}{16}, N_1Q_1 + N_2Q_2 \approx 2.83$
iv) Size ≤ 3, $p_1 = \frac{2}{4}, N_1 = 4, Q_1 = \frac{1}{2}, p_2 = 1, N_2 = 3, Q_2 = 0, N_1Q_1 + N_2Q_2 \approx 2$
v) Size ≤ 4 ,$p_1 = \frac{4}{6}, N_1 = 6, Q_1 = \frac{4}{9}, p_2 = 1, N_2 = 1, Q_2 = 0, N_1Q_1 + N_2Q_2 \approx 2.66$
The minimal weighted impurity measure is obtained by splitting on the Size ≤ 3.

Original

Spots = N

Poisonous	Size	Spots	Color
N	5	N	White
N	2	N	Brown
N	4	N	White
N	1	N	Brown

Spots = Y

Poisonous	Size	Spots	Color
N	2	Y	White
N	3	Y	Brown
Y	5	Y	White
Y	4	Y	Brown
Y	4	Y	Brown
Y	1	Y	White
Y	1	Y	Brown

Size $\leq$ 3

Poisonous	Size	Spots	Color
N	3	Y	Brown
N	2	Y	White
Y	1	Y	White
Y	1	Y	Brown

Size > 3

Poisonous	Size	Spots	Color
Y	5	Y	White
Y	4	Y	Brown
Y	4	Y	Brown

Let region 1, 2, 3 be {SPOT = N}, {SPOTS = Y, SIZE $\leq$ 3}, {SPOTS = N, SIZE > 3}
The predicted probability of poisonous is :

Region	Prob of Poisonous	Prob of not Poisonous
1	0 %	100 %
2	50 %	50 %
1	100 %	0 %

– 6.1.3.

In the given dataset, the three features are binary, there will be at most 8 nodes. If we build the tree until all nodes are either pure or cannot be split further, the error will occurs on the data that have the same feature values but different Y values. In the given dataset, the training error happens in :

Y	A	B	C
0	0	1	1
1	0	1	1

Y	A	B	C
0	1	1	1
1	1	1	1

There will be 2 sample incorrectly labeled, therefore, the training error is $\frac{2}{11} \approx 18.1\%$

- **6.2 Investigating Impurity Measures.**

– 6.2.1.

Misclassification rates:
Model A: $\frac{1}{4} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{2} = \frac{1}{4}$
Model B: $\frac{2}{6} \times \frac{6}{8} + 0 \times \frac{2}{8} = \frac{1}{4}$

Cross-entropy:
Model A: $-\frac{3}{4}log(\frac{3}{4}) \times \frac{1}{2} - \frac{1}{4}log(\frac{1}{4}) \times \frac{1}{2} \approx 0.122$
Model B: $-\frac{2}{6}log(\frac{2}{6}) \times \frac{6}{8} - 1log(1) \times \frac{2}{6} \approx 0.119$

Gini impurity:
Model A: $\frac{3}{4}\frac{1}{4}\frac{1}{2} + \frac{1}{4}\frac{3}{4}\frac{1}{2} = \frac{3}{16} = 0.1875$
Model B: $\frac{1}{3}\frac{2}{3}\frac{6}{8} + 0 = \frac{3}{16} \approx 0.1667$

Therefore, the Misclassification rates are identical for Model A and Model B, while the Cross-entropy and Gini impurity for Model B are less than that for Model A.

7 Representer Theorem

– 7.1.

$m_0 = Proj_M x$, and $||x||^2 = ||m_0||^2 + ||x - m_0||^2$
$||x|| = ||m_0|| \to ||x - m_0||^2 = 0$
$||x - m_0||^2 =< x - m_0, x - m_0 >= 0$ iff $x - m_0 = \vec{0}$ by positive-definiteness of inner product.
Therefore $||x|| = ||m_0||$ only when $x = m_0$

– 7.2.

$R(\cdot)$ is strictly increasing, let M = $span(\psi(x_1), ..., \psi(x_n))$, assume w^* is a minimizer, and w = Proj$_M w^*$.
So $\exists \alpha$ s.t. $w = \sum \alpha_i \psi(x_i)$

case 1: $||w|| = ||w^*||$
from 7.1, we know that if $||w|| = ||w^*||$, then $x = m_0$
This immediately shows that w is a minimizer and w has the form $\sum \alpha_i \psi(x_i)$

case 2: $||w|| < ||w^*||$
Since $R(\cdot)$ is strictly increasing, $R(||w||) < R(||w^*||)$
We know that $w^\perp = w^* - w$ is orthogonal to M

$$< w^*, \psi(x_i) >=< w + w^\perp, \psi(x_i) >=< w, \psi(x_i) > \tag{12}$$

$$L(< w^*, \psi(x_1) >, ..., < w^*, \psi(x_n) >) = L(< w, \psi(x_1) >, ..., < w, \psi(x_n) >) \tag{13}$$

Therefore,

$$J(w) = R(||w||) + L(< w, \psi(x_1) >, ..., < w, \psi(x_n) >) \tag{14}$$

$$< R(||w^*||) + L(< w^*, \psi(x_1) >, ..., < w^*, \psi(x_n) >) = J(w^*) \tag{15}$$

This contradict to the fact that w^* is a minimizer. Therefore, this case is discarded.
In conclusion, only case 1 is possible, then we proved that all minimizers have the form $w = \sum \alpha_i \psi(x_i)$

- **7.3.**

$w \in \Re^d$, let $A = \begin{pmatrix} \psi_1(x_1) & \cdots & \psi_d(x_1) \\ \vdots & \vdots & \vdots \\ \psi_1(x_n) & \cdots & \psi_d(x_n) \end{pmatrix}$ be the design matrix, b be the bias,$L(w) = Aw + b$

R and L are both convex, let $w_1, w_2 \in \Re^d$ and $0 \leq c \leq 1$

$$J(cw_1 + (1-c)w_2) = R(||cw_1 + (1-c)w_2||) + L(A(cw_1 + (1-c)w_2) + b) \tag{16}$$

Since L is convex and $Aw + b$ is an affine function, $L(Aw + b)$ is convex.

$$L(A(cw_1 + (1-c)w_2) + b) \leq cL(Aw_1 + b) + (1-c)L(Aw_2 + b) \tag{17}$$

$||cw_1 + (1-c)w_2|| \leq c||w_1|| + (1-c)||w_2||$, R is increasing and convex

$$R(||cw_1 + (1-c)w_2||) \leq R(c||w_1|| + (1-c)||w_2||) \leq cR(||w_1||) + (1-c)R(||w_2||) \tag{18}$$

eqn (17) and (18) together shows that

$$J(cw_1 + (1-c)w_2) \leq cJ(w_1) + (1-c)J(w_2) \tag{19}$$

This proves J is convex.

8 Ivanov and Tikhonov Regularization

• 8.1 Tikhnov optimal implies Ivanov optimal.

- **8.1.1.**

Since for some $\lambda > 0, f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ \phi(f) + \lambda\Omega(f)$, then :

$$\nabla\phi(f^*) + \lambda\nabla\Omega(f^*) = 0 \tag{20}$$

Let $r = \Omega(f^*) > 0$, we need to show that :
f^* is also a solution to $\underset{f \in \mathcal{F}}{\operatorname{argmin}}\ \phi(f)$ s.t $\Omega(f) \leq \Omega(f^*)$
The Lagrangian to this Ivanov problem is :

$$L(f) = \phi(f) + \lambda(\Omega(f) - \Omega(f^*)) \tag{21}$$

We claim that f^* is a solution, by the first order condition :

$$\nabla L(f^*) = \nabla\phi(f^*) + \lambda\nabla\Omega(f^*) = 0 \qquad \text{(from (20))} \tag{22}$$

$$\Omega(f^*) - \omega(f^*) = 0 \tag{23}$$

Therefore f^* is also a Ivanov solution

• 8.2 Ivanov optimal implies Tikhonov optimal .

- **8.2.1.**

The Lagrangian for Ivanov problem is :

$$L(w, \lambda) = \phi(w) + \lambda(\Omega(w) - r) \tag{24}$$

– **8.2.2.**

The duel problem is :

$$g(\lambda) = \inf_{w} L(w, \lambda) = \min_{w} \ \phi(w) + \lambda(\Omega(w) - r) \tag{25}$$