Assignment 4

Benjamin Jakubowski

March 21, 2016

2. Positive Semidefinite Matrices

2.1 Orthogonal, non-symmetric matrix

Let

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Then the columns of A are clearly orthonormal, but $A \neq A^T$ since

$$A^T = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

2.2 Positive semidefinite matrix has non-negative eigenvalues

Let M be a positive semidefinite matrix. Then, by the definition of positive semidefinite, m is a real symmetric matrix. Thus, by the spectral theorem, we can diagonalize M as

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

Then, since Q is orthogonal (columns are orthonormal eigenvectors), we have

$$Q^T M Q = Q^T Q \Sigma Q^T Q = \mathbb{I}_n \Sigma \mathbb{I}_n = \Sigma$$

Now,

$$\Sigma = Q^T M Q = \begin{pmatrix} q_1^T M q_1 & q_1^T M q_2 & \cdots & q_1^T M q_n \\ q_2^T M q_1 & q_2^T M q_2 & \cdots & q_2^T M q_n \\ \vdots & \vdots & \cdots & \vdots \\ q_n^T M q_1 & q_n^T M q_2 & \cdots & q_n^T M q_n \end{pmatrix}$$

Finally, since $q_i^T M q_i \geq 0$ (by M positive semidefinite), this implies all of the eigenvalues of M (i.e. the diagonal entries of Σ) are non-negative.

2.3 $M = BB^T$ if and only if M is psd

We first prove the forward direction: if M is positive semidefinite, then (by above) we know its eigenvalues are non-negative. Thus, (noting Σ is diagonal with non-negative entries) we can rewrite

$$\Sigma = \Sigma'^2 = \Sigma'^T \Sigma'$$

where

$$\Sigma' = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{pmatrix}$$

Thus,

$$M = Q^T \Sigma Q = Q^T \Sigma'^T \Sigma' Q = (\Sigma' Q)^T \Sigma' Q$$

so setting $B = (\Sigma' Q)^T$ yields the desired results: $M = BB^T$.

Next, we prove the backwards direction: if $M = BB^T$ for some matrix B, then

$$x^T M x = x^T B B^T x = (B^T x)^T (B^T x) \geq 0$$

since for any vector $v, v^T v \geq 0$. Thus we have

$$x^T M x \geq 0$$

so M is positive semidefinite. Having demonstrated both directions, we conclude a symmetric matrix M can be expressed as $M = BB^T$ for some matrix B if and only if M is positive semidefinite.

3. Positive Definite Matrices

3.1 Eigenvalues of Positive Definite Matrix are Positive

Let M be a positive definite matrix. Then (similarly to 2.2) we have

$$\Sigma = Q^T M Q = \begin{pmatrix} q_1^T M q_1 & q_1^T M q_2 & \cdots & q_1^T M q_n \\ q_2^T M q_1 & q_2^T M q_2 & \cdots & q_2^T M q_n \\ \vdots & \vdots & \cdots & \vdots \\ q_n^T M q_1 & q_n^T M q_2 & \cdots & q_n^T M q_n \end{pmatrix}$$

Finally, since $q_i^T M q_i > 0$ (by M positive definite), this implies all of the eigenvalues of M (i.e. the diagonal entries of Σ) are positive.

3.2 $Q\Sigma^{-1}Q^T$ is inverse of M

First, let M be a symmetric positive definite matrix. Then the spectral theorem gives:

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

where Q is the matrix with orthonormal eigenvectors of M as columns. Since Q is orthonormal, we know $Q^T = Q^{-1}$. Thus,

$$(Q\Sigma^{-1}Q^T)M = Q\Sigma^{-1}Q^T Q\Sigma Q^T = Q\Sigma^{-1}\mathbb{I}_n\Sigma Q^T = Q\Sigma^{-1}\Sigma Q^T = Q\mathbb{I}_n Q^T = QQ^T = \mathbb{I}_n$$

so $Q\Sigma^{-1}Q^T$ is the inverse of M.

3.3 $M + \lambda\mathbb{I}$ is symmetric positive definite for any $\lambda > 0$

If M is a positive semidefinite matrix then recall

- $M = Q\Sigma Q^T$
- M has eigenvalues $\lambda_i \geq 0$

Now consider $M + \lambda\mathbb{I}$ for some $\lambda > 0$. Before we proceed, note

$$\lambda\mathbb{I} = \lambda QQ^T = Q\lambda\mathbb{I}Q^T$$

Thus,

$$M + \lambda\mathbb{I} = Q\Sigma Q^T + Q\lambda\mathbb{I}Q^T = Q(\Sigma + \lambda\mathbb{I})Q^T$$

But then the eigenvalues of $M + \lambda\mathbb{I}$ are the diagonal entries of $\Sigma + \lambda\mathbb{I}$, which are all positive. Therefore, since $M + \lambda\mathbb{I}$ is symmetric matrix with positive eigenvalues, it is positive definite. It's inverse is then (by 3.2)

$$(M + \lambda\mathbb{I})^{-1} = Q^T(\Sigma + \lambda\mathbb{I})^{-1}Q$$

(noting $(\Sigma + \lambda\mathbb{I})^{-1}$ is defined since the matrix is diagonal with non-zero entries).

3.4 $M + N$ positive definite

Let M be positive semidefinite, and N be positive definite. Now consider $M + N$. Well,

$$x(M + N)x^T = \underbrace{xMx^T}_{\geq 0} + \underbrace{xNx^T}_{>0} > 0$$

So $M + N$ is positive definite.

4. Kernel Matrices

4.1 Kernel matrix as pairwise distance and vector length

Let X be the matrix whose rows are the vectors in a set $S = \{x_1, \cdots, x_m\}$, and let $K = XX^T$. Then note

$$K = XX^T \begin{pmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & \cdots & \langle x_1, x_n \rangle \\ \langle x_2, x_1 \rangle & \langle x_2, x_2 \rangle & \cdots & \langle x_2, x_n \rangle \\ \vdots & \vdots & \cdots & \vdots \\ \langle x_n, x_1 \rangle & \langle x_n, x_2 \rangle & \cdots & \langle x_n, x_n \rangle \end{pmatrix}$$

Now, note the distance between any two vectors x_i, x_j in S is

$$\begin{aligned} d(x_i, x_j) &= ||x_i - x_j|| \\ &= \sqrt{\langle x_i - x_j, x_i - x_j \rangle} \\ &= \sqrt{\langle x_i, x_i \rangle - 2\langle x_i, x_j \rangle + \langle x_j, x_j \rangle} \end{aligned}$$

This can clearly be computed from K. Similarly, the length of any vector x_i in S is just $\sqrt{\langle x_i, x_i \rangle}$, which is just the square root of the i, i entry of K. Thus, knowing K implies you know the set of pairwise distances and the vector lengths.

The reverse implication follows directly. If you know the pairwise distances and the vector lengths, then (for arbitrary $x_i, x_j \in S$) note

$$(d(x_i, x_j))^2 = \langle x_i, x_i \rangle - 2\langle x_i, x_j \rangle + \langle x_j, x_j \rangle$$

Thus, subtracting off the squared lengths of x_i and x_j and dividing by -2 yields $\langle x_i, x_j \rangle$. Thus, we know both the diagonal and off-diagonal entries and can reconstruct K.

5. Kernel Ridge Regression

5.1 Minimizing the ridge objective

Recall the ridge objective function is

$$J(w) = ||Xw - y||^2 + \lambda ||w||^2$$

Let $w*$ be a minimizer of $J(w)$. Then

$$\begin{aligned} & \nabla_w J(w*) = 2X^T(Xw^* - y) + 2\lambda w^* = 0 \\ \implies \quad & X^T X w^* + \lambda \mathbb{I} w^* = X^T y \end{aligned}$$

Thus, the minimizer is

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

Note $X^T X + \lambda \mathbb{I}$ is invertible, since $X^T X$ is positive semidefinite (by 2.3), so we can apply the result of 3.3.

5.2 Rewriting expression from 5.1

$$\begin{aligned}
X^TXw^* + \lambda\mathbb{I}w^* &= X^Ty \\
\implies \qquad \lambda\mathbb{I}w^* &= X^Ty - X^TXw^* \\
\implies \qquad \mathbb{I}w^* &= \frac{1}{\lambda}(X^Ty - X^TXw^*) \\
\implies \qquad w^* &= \frac{1}{\lambda}(X^Ty - X^TXw^*)
\end{aligned}$$

Now we find α such that $w^* = X^T\alpha$. Factoring out X^T from the expression for w^* yields

$$w^* = X^T\frac{y - Xw^*}{\lambda}$$

Thus, setting $\alpha = \frac{y-Xw^*}{\lambda}$ we get $w^* = X^T\alpha$. Note, by way of interpretation, we see $\alpha = \frac{r}{\lambda}$, where r is the residual vector.

5.3 Rewriting expression from 5.3

w^* is in the span of the data since

$$w^* = X^T\alpha = \sum_{i=1}^{n} \alpha_i x_i^T$$

(to see this, recall the matrix-vector product Mv is just a linear combination of the columns of M, weighted by coefficients $v[i]$).

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

Starting with the expression,

$$\alpha = \frac{y - Xw^*}{\lambda}$$

we substitute in our expression for $w^* = X^T\alpha$ to get

$$\begin{aligned}
\alpha &= \frac{y - XX^T\alpha}{\lambda} \\
&= \frac{1}{\lambda}y - \frac{1}{\lambda}XX^T\alpha \\
\implies \qquad \lambda\alpha + XX^T\alpha &= y \\
(\lambda\mathbb{I} + \underbrace{XX^T}_{\text{psd}})\alpha &= y \\
\alpha &= (\lambda\mathbb{I} + XX^T)^{-1}y
\end{aligned}$$

Note the inverse in the final expression exists (by 3.3 and XX^T positive semidefinite, which follows from 2.3).

5.5 Kernelized expression for Xw^*

To find a kernelized expression for Xw^*, we first substitute in for w^*, then for α:

$$\begin{aligned} Xw^* &= XX^T\alpha \\ &= XX^T(\lambda\mathbb{I} + XX^T)^{-1}y \end{aligned}$$

Since the data only appears in terms of the kernel matrix $K = XX^T$, we're done.

5.6 Kernelized expression for the prediction $f(x) = x^T w^*$

Again, we use a similar approach as 5.5 (to ease notation, we call our new point z):

$$\begin{aligned} f(z) = z^T w^* &= z^T X^T \alpha \\ &= z^T X^T(\lambda\mathbb{I} + XX^T)^{-1}y \\ &= k_z^T(\lambda\mathbb{I} + XX^T)^{-1}y \end{aligned}$$

where

$$k_z = \begin{pmatrix} z^T x_1 \\ \vdots \\ z^T x_n \end{pmatrix}$$

Since the new point z only appears in inner products with other x's, we're done.

6. Decision Trees

6.1 Building Trees by Hand

To build our binary classification tree on this data using the Gini index, we need to consider six splits: splitting on *spots*, *color*, and on *size* $\leq 1, 2, 3,$ or 4. Results for these splits are shown below:

Split	$\hat{p}_1$	Q_1	N_1	$\hat{p}_2$	Q_2	N_2	$N_1Q_1 + N_2Q_2$
spots	0.714	0.408	7	0	0	4	2.857
color	0.4	0.48	5	0.5	0.5	6	5.4
size ≤ 1	0.667	0.444	3	0.375	0.469	8	5.083
size ≤ 2	0.4	0.48	5	0.5	0.5	6	5.4
size ≤ 3	0.333	0.444	6	0.6	0.48	5	5.067
size ≤ 4	0.444	0.494	9	0.5	0.05	2	5.444

Thus, the first split (using the Gini index) is on *spots* (Y/N).

6.2 Splitting next two parts

After splitting on *spots*, note the $spots = No$ node is pure (entirely non-poisonous), so no additional splits are needed. On the other side of the stump, we continue splitting:

Split	$\hat{p}_1$	Q_1	N_1	$\hat{p}_2$	Q_2	N_2	$N_1Q_1 + N_2Q_2$
color	0.666	0.444	3	0.75	0.375	4	2.833
$size \leq 1$	1	0	2	0.6	0.48	5	2.4
$size \leq 2$	0.666	0.444	3	0.75	0.375	4	2.833
$size \leq 3$	0.5	0.5	4	1	0	3	2
$size \leq 4$	0.666	0.444	6	1	0	1	2.667

Thus, we split on $spots \leq 3$. Thus, for our decision function, we've partitioned the input space into three regions. These regions, the number of observations, and the predicted probability of poisonous are given below:

Region	Number of observations	Probability of poisonous
No spots	4	0
Spots and size ≤ 3	4	0.5
Spots and size > 3	3	1

6.3 Maximally deep binary tree

Recall the data have three binary features: A, B, and C, and we build the tree so all terminal nodes are either pure or cannot be split further. In this case, "cannot be split further" implies we have fully specified the values of A, B, and C (ex: $A = 0, B = 1$, and $C = 1$), but the values of the target in this region are not pure.

Given this approach (and the given data), we would misclassify two instances out of 11. Thus, our training error would be 18.2%.

6.2 Investigation Impurity Measures

In this problem, we consider two decision stumps: A and B. The stumps partition the input space into two regions, with the following target class distributions:

Stump	R_1 distribution	R_2 distribuiton
A	(300, 100)	(100, 300)
B	(200, 400)	(200, 0)

To compare these splits, the misclassification rate, cross-entropy, and Gini impurity are given below:

Stump	Metric	Value
A	Misclassification	$1/2 \cdot 1/4 + 1/2 \cdot 1/4 = 0.25$
B	Misclassification	$3/4 \cdot 1/3 + 1/4 \cdot 0 = 0.25$
A	Cross entropy	$2 \cdot (-1/2 \cdot (0.25 \cdot \log 0.25 + 0.75 \cdot \log 0.75)) = 0.56$
B	Cross entropy	$-3/4 \cdot (1/3 \cdot \log 1/3 + 2/3 \cdot \log 2/3) - 1/4 \cdot (1 \cdot \log 1 + 0 \cdot \log 0) = 0.477$
A	Gini	$2 \cdot (1/2 \cdot (2 \cdot 0.25 \cdot (1 - 0.75))) = 0.375$
B	Gini	$3/4 \cdot (2 \cdot 1/3 \cdot (1 - 1/3)) + 1/4 \cdot (2 \cdot 1 \cdot 0) = 0.333$

Thus, A and B have the same misclassification rates, but B has both a lower cross entropy and a lower Gini impurity measure.

7. Representer Theorem

7.1 $||m_0|| = ||x||$ only if $m_0 = x$

Recall M is a closed subspace of a Hilbert space $\mathcal{H}$, and $m_0 = \text{Proj}_M x$ for some x in $\mathcal{H}$. We want to show that we have $||m_0|| = ||x||$ only if $m_0 = x$.
Well, by the Pythagorean Theorem, we know

$$||x||^2 = ||m_0||^2 + ||x - m_0||^2$$

Thus, $||x||^2 = ||m_0||^2$ only if $||x - m_0||^2 = 0$. But then, note

$$||x - m_0||^2 = \langle x - m_0, x - m_0 \rangle = 0 \iff x - m_0 = 0$$

since inner products are positive definite. Thus, we know $||m_0|| = ||x||$ only if $m_0 = x$.

7.2 Completed proof of Representer Theorem

Now we complete the proof of the Representer Theorem. Recall the set up of the theorem: Let

$$J(w) = R(||w||) + L(\langle w, \psi(x_1) \rangle, \cdots, \langle w, \psi(x_n) \rangle)$$

where

- $R : \mathbb{R}^{\geq 0} \to \mathbb{R}$ is a nondecreasing regularization term
- $L : \mathbb{R}^n \to \mathbb{R}$ is an arbitrary loss term

We already showed if $J(w)$ has a minimizer, then it has a minimizer of the form

$$w^* = \sum_{i=1}^{n} \alpha_i \psi(x_i)$$

We complete the proof by showing if R (the regularization term) is strictly increasing, then all minimizers have this form.

First, let w^* be a minimizer, and let $M = \text{span}(\psi(x_1), \cdots \psi(x_n))$. Now, let $w = \text{Proj}_M w^*$, so there exists an α such that $w = \sum_{i=1}^n \alpha_i \psi(x_i)$.
Then, $w^\perp := w^* - w$ is orthogonal to M. Moreover, since projections decrease norms, we know $||w|| \leq ||w^*||$.
Now we consider two cases:

- **Case 1:** Assume $||w|| = ||w^*||$. Then (by the result of 7.1), $w = w^*$, so w is of the form $w = \sum_{i=1}^n \alpha_i \psi(x_i)$.
- **Case 2:** Otherwise, $||w|| < ||w^*||$. Then, since R is strictly increasing, $R(||w||) < R(||w^*||)$. Now, let's compare the loss terms: for all $i \in 1 \cdots n$,

  $$\langle w^*, \psi(x_i)\rangle = \langle w + w^\perp, \psi(x_i)\rangle = \langle w, \psi(x_i)\rangle$$

  so
  $$L(\langle w^*, \psi(x_1)\rangle, \cdots, \langle w^*, \psi(x_n)\rangle) = L(\langle w, \psi(x_1)\rangle, \cdots, \langle w, \psi(x_n)\rangle)$$

  But this implies

  $$\begin{aligned} J(w) &= R(||w||) + L(\langle w, \psi(x_1)\rangle, \cdots, \langle w, \psi(x_n)\rangle) \\ &< R(||w^*||) + L(\langle w^*, \psi(x_1)\rangle, \cdots, \langle w^*, \psi(x_n)\rangle) = J(w^*) \end{aligned}$$

  This yields a contradiction (since it implies w^* is not a minimizer).

Thus, we cannot have **Case 2**, so in all cases

$$w^* = \sum_{i=1}^n \alpha_i \psi(x_i)$$

7.3 Case where R and L are both convex in their arguments

Now suppose R and L are both convex in their arguments. We show this implies J is a convex function of w, and as such J has a minimizer of the form $w^* = \sum_{i=1}^n \alpha_i \psi(x_i)$.
First, let's consider R. Recall R is non-decreasing. Then, by the convexity of norms, we know

$$||\theta w + (1-\theta)w'|| \leq \theta||w|| + (1-\theta)||w'||$$

Thus (since R is non-decreasing)

$$R(||\theta w + (1-\theta)w'||) \leq R(\theta||w|| + (1-\theta)||w'||)$$

Then, since R is convex in it's inputs

$$R(\theta||w|| + (1-\theta)||w'||) \leq \theta R(||w||) + (1-\theta)R(||w'||)$$

Thus R is convex in terms of w.
Now let's consider L. First, note L is a function of $(\langle w, \psi(x_1)\rangle, \cdots, \langle w, \psi(x_n)\rangle)$.

However, we can represent

$$(\langle w, \psi(x_1)\rangle, \cdots, \langle w, \psi(x_n)\rangle) = \Psi A w$$

since every inner product in R^d can be represented as $\langle x, y\rangle = x^t A y$ for some A[1].
Thus, since L is convex and $\Psi A w$ is affine, their composition is convex, so $L(\Psi A w)$ is convex in w.
Now we're ready to tackle $J(w)$. Since

$$J(w) = R(||w||) + L(\Psi A w)$$

$R(||w||)$ and $L(\Psi A w)$ are both convex in w, and positive weighted sums of convex functions are convex, we conclude J is convex in w.

8. Ivanov and Tikhonov Regularization

8.1 Tikhonov optimal implies Ivanov optimal

Suppose for some $\lambda > 0$ we have the Tikhonov regularization solution

$$f_* = \arg\min_{f \in \mathcal{F}} [\phi(f) + \lambda \Omega(f)]$$

We show f_* is the Ivanov solution by finding an $r > 0$ such that

$$f_* = \arg\min_{f \in \mathcal{F}} \phi(f) \text{ subject to } + \Omega(f) \leq r$$

Specifically, take $r = \Omega(f*)$. Then we know $\Omega(f_*) \leq r = \Omega(f_*)$, so we know f_* is in the feasibility set for the Ivanov problem. We proceed using contradiction to show it is in fact optimal: Assume that f_* is not the Ivanov solution. Then there is a solution $f' \neq f_*$ such that $\phi(f') < \phi(f_*)$. Moreover, since f' is the solution, it is feasible, so $\Omega(f') \leq r = \Omega(f_*)$. Additionally (multiplying both sides by $\lambda > 0$) we have $\lambda\Omega(f') \leq r = \lambda\Omega(f_*)$
Thus, adding the two inequalities yields

$$\phi(f') + \lambda\Omega(f') < \phi(f_*) + \lambda\Omega(f_*)$$

This is our contradiction, since it implies f_* is not the Tikhonov solution. Thus, f_* is the Ivanov solution.

8.2 Ivanov optimal implies Tikhonov optimal

Now assume $\mathcal{F} = \{f_w(x) : \mathcal{X} \to \mathbb{R} | w \in \mathbb{R}^d\}$. Let w^* be a solution to the Ivanov problem. We proceed assuming strong duality holds for this problem, and the dual solution is attained, and show there exists a $\lambda \geq 0$ such that $w^* = \arg\min_{w \in \mathbb{R}^d}[\phi(w) + \lambda\Omega(w)]$.

[1] See `https://math.berkeley.edu/~peyam/Math110Sp13/Handouts/Dot%20products.pdf`

8.2.1 Langrangian

The langrangian for the Ivanov optimization problem is

$$L(w, \lambda) = \phi(w) + \lambda(\Omega(w) - r)$$

8.2.2 Dual optimization problem

The primal problem is

$$p^* = \inf_{w \in \mathbb{R}^d} \sup_{\lambda \geq 0} \phi(w) + \lambda(\Omega(w) - r)$$

The dual problem is

$$d^* = \sup_{\lambda \geq 0} \left[\inf_{w \in \mathbb{R}^d} \phi(w) + \lambda(\Omega(w) - r) \right]$$

Letting $g(\lambda) = \inf_{w \in \mathbb{R}^d} \phi(w) + \lambda(\Omega(w) - r)$ yields

$$d^* = \sup_{\lambda \geq 0} g(\lambda)$$

8.2.3 Minimum in $g(\lambda^*)$ attained at w^*

Now recall we assumed the dual solution is attained, so let $\lambda^* = \arg\max_{\lambda \geq 0} g(\lambda)$. We show the minimum in $g(\lambda^*)$ attained at w^*. First, by strong duality, we have

$$\phi(w^*) = g(\lambda^*)$$

Expanding we find

$$\begin{aligned} \phi(w^*) &= g(\lambda^*) \\ &= \inf_{w \in \mathbb{R}^d} \phi(w) + \lambda^*(\Omega(w) - r) \\ &\leq \phi(w^*) + \lambda^*(\Omega(w^*) - r) \end{aligned}$$

Now, since w^* is optimal (the solution), it is feasible, so $\Omega(w^*) - r \leq 0$. Thus (recalling $\lambda \geq 0$), we have

$$\phi(w^*) \leq \phi(w^*) + \underbrace{\lambda^*(\Omega(w^*) - r)}_{\leq 0}$$

which implies $\lambda^*(\Omega(w^*) - r) = 0$.
Hence, we end with

$$\phi(w^*) = g(\lambda^*) = \inf_{w \in \mathbb{R}^d} \phi(w) + \lambda^*(\Omega(w) - r) \leq \phi(w^*) + \lambda^*(\Omega(w^*) - r) = \phi(w^*)$$

so we conclude the minimum of the expression in $g(\lambda^*)$ is attained at w^*.

Finally, we conclude by showing

$$w^* = \arg\min_{w\in\mathbb{R}^d}[\phi(w) + \lambda^*\Omega(w)]$$

From above, we know (relaxing our rigor and using arg min instead of inf)

$$\begin{aligned} w^* &= \arg\min[\phi(w) + \lambda^*(\Omega(w) - r)] \\ &= \arg\min[\phi(w) + \lambda^*\Omega(w) - \lambda^* r] \\ &= \arg\min[\phi(w) + \lambda^*\Omega(w)] \qquad \text{(dropping the constant } -\lambda^* r) \end{aligned}$$

8.3 Ivanov implies Tikhonov for Ridge Regression

To show that Ivanov implies Tikhonov for the ridge regression problem (square loss with ℓ_2 regularization), we need to demonstrate strong duality and that the dual optimum is attained. Both of these things are implied by Slater?s constraint qualifications. To demonstrate Slater's constraint qualifications are met, we need to find a strictly feasible point.
First, recall the Ivanov form of the ridge regression problem is

$$\min_{w\in\mathbb{R}^d} \frac{1}{n}||Xw - y||^2 \text{ s.t. } ||w||^2 \leq r$$

for some $r > 0$. Then note $w = 0$ is strictly feasible (since $r > 0$), and the objective and constraints are convex, so by Slater's constraint qualifications strong duality holds and the dual optimum is attained.

9. Novelty Detection

9.1 Formulating the algorithm as an optimization problem

Recall a novelty detection algorithm can be based on an algorithm that finds the smallest possible sphere containing the data in feature space- our goal is to formulate this algorithm as an optimization problem. Well, this problem is just

$$\min_{b\in\mathbb{R}} b$$

$$\text{s.t.}||\phi(x_i)||_2 \leq b \text{ for all } i \in \{1, \cdots, n\}$$

9.2 The Lagrangian and the primal problem

The Lagrangian for this problem is

$$L(b, \lambda) = b + \sum_{i=1}^{n} \lambda_i(||\phi(x_i)||_2 - b)$$

The primal problem is

$$p^* = \inf_{b\in\mathbb{R}} \sup_{\lambda\succeq 0} \left[b + \sum_{i=1}^{n} \lambda_i (||\phi(x_i)||_2 - b) \right]$$

9.3 Demonstrating strong duality

We have strong duality by Slater's conditions:

- Convex objective
- Convex (affine) constraints (note $||\phi(x_i)||$ is constant given data)
- Feasibility by $b = \max\{\phi(x_1), \cdots, \phi(x_n)\}$

Thus the primal and dual optima are equal.

9.4 Solve the inner minimization and give the dual problem

First, the dual problem is

$$d^* = \sup_{\lambda\succeq 0} \inf_{b\in\mathbb{R}} \left[b + \sum_{i=1}^{n} \lambda_i (||\phi(x_i)||_2 - b) \right]$$

We now solve the inner minimization problem- note L is differentiable with respect to b, so we find $\inf_{b\in\mathbb{R}}$ by solving $\frac{\partial}{\partial b}L = 0$:

$$\begin{aligned} \frac{\partial}{\partial b}L &= \frac{\partial}{\partial b}\left[b + \sum_{i=1}^{n} \lambda_i (||\phi(x_i)||_2 - b) \right] \\ &= 1 - \sum_{i=1}^{n} \lambda_i \triangleq 0 \\ \implies \qquad \sum_{i=1}^{n} \lambda_i &= 1 \end{aligned}$$

Now let's return to the dual problem: note

$$\begin{aligned} d^* &= \sup_{\lambda\succeq 0} \inf_{b\in\mathbb{R}} \left[b + \sum_{i=1}^{n} \lambda_i (||\phi(x_i)||_2 - b) \right] \\ &= \sup_{\lambda\succeq 0} \inf_{b\in\mathbb{R}} \left[b - b\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} \lambda_i ||\phi(x_i)||_2 \right] \end{aligned}$$

Then, at the minimum (since $\sum_{i=1}^{n} \lambda_i = 1$), this yields

$$
\begin{aligned}
d^* = \max_{\lambda} & \sum_{i=1}^{n} \lambda_i ||\phi(x_i)||_2 \\
\text{s.t.} & \sum_{i=1}^{n} \lambda_i = 1 \\
& \lambda_i \geq 0 \text{ for all } i \in \{1, \cdots, n\}
\end{aligned}
$$

9.5 Expression for optimal sphere

First, recall (by strong duality) $d^* = w^*$. Thus, our optimal sphere is clearly

$$d^* = \max\{||\phi(x_1)||_2, \cdots, ||\phi(x_n)||_2\}$$

(noting this point is clearly feasible and any other point in the convex hull of $\{||\phi(x_1)||_2, \cdots, ||\phi(x_n)||_2\}$ has a lower value).

9.6 Complementary slackness conditions

The complementary slackness conditions for this problem are

$$\lambda_i(||\phi(x_i)||_2 - b) = 0 \text{ for all } i \in \{1, \cdots, n\}$$

Thus, the "support vectors" are those x_i where $||\phi(x_i)||_2 = b$.

9.7 Applying this algorithm to detect "novel"[2] instances

To apply this algorithm to detect "novel" instances, I would

1. Sort the data by decreasing length in feature space (i.e. by $||\phi(x_i)||_2$)
2. Starting with the first (greatest length) observation, for the i^{th} observation
   a) Determine the smallest possible sphere that contains the data (i.e. b_i)
   b) Remove the observation from the dataset
3. Compare the measured b_is. For some threshold $\alpha > 0$ label an observation "novel" if
   - **Case 1**: $i = 1$, and $b_1 - b_2 > \alpha$.
   - **Case 2**: $1 < i < n$, $b_i - b_{i+1} > \alpha$, and $b_{i-1} - b_i > \alpha$.
   - **Case 3**: $i = n$, $b_{n-1} - b_n > \alpha$.

[2] Note here I am essentially defining novel instances as those instances that at least α from their nearest neighbors. This algorithm would fail if two instances were near in feature space, but both anomalous