COMPSCI 671 **Name:** *David Hunt*
HW 3
15 November, 2022

Agreement 1) This assignment represents my own work. I did not work on this assignment with others. All coding was done by myself.

Agreement 2) I understand that if I struggle with this assignment that I will reevaluate whether this is the correct class for me to take. I understand that the homework only gets harder.

1 Hoeffding and Beyond

1.1 Markov's Inequality

For any non-negative random variable X, we seek to prove the following:
$P(X \geq \epsilon) \leq \frac{E(X)}{\epsilon}$

for any positive real number $\epsilon > 0$

1. Let X be a random variable such that $Pr(X \geq 0) = 1$, and let the probability function of X be $f(x)$.
2. Let $E(X) \triangleq$ mean of X
3. Using the definition of the mean, we can say the following
   $E(X) = \sum_x xf(x) = \sum_{x<t} xf(x) = \sum_{x \geq t} xf(x)$
4. From this, we can say the following:
   $\Rightarrow E(X) \geq \sum_{x \geq t} xf(x) \geq \sum_{x \geq t} tf(x) = tPr(X \geq t)$
5. From this, we can arrive at the final proof statement
   $\Rightarrow Pr(X \geq t) = \frac{E(X)}{t}$

1.2 Chebyschev's Inequality

We seek to prove Chebyschev's inequality which states that for any random variable X with mean μ and variance σ^2, we have:
$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$

1. Let us define a random variable $Y = |X - E(X)|^2$ where $E(x) = \mu \triangleq$ mean of X
2. by definition of variance we can say the following:
   $E(Y) = E(|X - E(X)|^2) = Var(X) = \sigma^2$
3. Applying Markov's inequality to Y, we can say the following
   $P(|X - \mu| \leq \epsilon) = P(|X - \mu|^2 \leq \epsilon^2) = P(Y \leq \epsilon^2) \leq \frac{E(Y)}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}$

1.3 Polynomial Markov's Inequality

For a random variable X with a mean μ and kth central moment $E[|X-\mu|^k]$, we seek to prove the polynomial version of Markov's inequality which states that:
$P(|X-\mu| \geq \epsilon) \leq \frac{E[|X-\mu|^k]}{\epsilon^k}$

1. Define $Y = |X-\mu|^k$
2. $\Rightarrow E(Y) = E[|X-\mu|^k$
3. From here, we can then apply Markov's Inequality as proved in part 1 to complete the proof:
   $P(|X-\mu| \geq \epsilon) = P(|X-\mu|^k \geq \epsilon^k) = P(Y \geq \epsilon^k) \leq \frac{E(Y)}{\epsilon^k} = \frac{E[|X-\mu|^k}{\epsilon^k}$

1.4 Chernoff Bound

Prove the following:
$P(X-\mu \geq \epsilon) \leq \inf_{\lambda \geq 0} \frac{M_{X-\mu}(\lambda)}{exp(\lambda\epsilon)}$
where μ is the mean of $X, \lambda > 0$ is a positive real number and ϵ is any real number

1. Start by defining a random variable $Y = X - \mu$
2. Next, we can say: $P(X-\mu \geq \epsilon) = P(Y \geq \epsilon) = P(exp(\lambda Y) \geq exp(\lambda\epsilon))$
3. Apply Markov's inequality:
   $P(exp(\lambda Y) \geq exp(\lambda\epsilon)) \leq \frac{E(exp(\lambda Y))}{exp(\lambda\epsilon)} = \frac{M_Y(\lambda)}{exp(\lambda\epsilon)}$
4. Since this holds for all λ it also holds for the infimum:
   $P(exp(\lambda Y) \geq exp(\lambda\epsilon)) \leq \inf_{\lambda \geq 0} \frac{M_Y(\lambda)}{exp(\lambda\epsilon)}$
5. Simplifying and substituting arrives at the final form of the proof
   $P(X-\mu \geq \epsilon) = P(Y \geq \epsilon) = P(exp(\lambda Y) \geq exp(\lambda\epsilon)) \leq \inf_{\lambda \geq 0} \frac{M_Y(\lambda)}{exp(\lambda\epsilon)} = \inf_{\lambda \geq 0} \frac{M_{X-\mu}(\lambda)}{exp(\lambda\epsilon)}$

1.5 Hoeffding's inequality

Fist, Hoeffding's Lema is defined as follows for any random variable X taking values in the interval $[a, b]$:
$E[exp(\lambda(X-\mu))] \leq exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$ Given $X_1..., X_n$ as IID bounded random variables in the range $[a, b]$ with a common mean μ, for any positive integer $n > 0$ and any real number ϵ, prove the following inequality which is a one-sided version of Hoeffding's Inequality:
$P\left(\left(\frac{1}{n}\sum_{i=1}^n X_i\right) - \mu \geq \epsilon\right) \leq exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)$

1. Start by multiplying both sides of the probability function by n and simplifying:
   $P\left(\left(\frac{1}{n}\sum_{i=1}^n X_i\right) - \mu \geq \epsilon\right) = P\left(\left(\sum_{i=1}^n X_i\right) - n\mu \geq n\epsilon\right) = P\left(\left(\sum_{i=1}^n X_i - \mu\right) \geq n\epsilon\right)$

2. Apply the Chernoff Bound:
   $P\left(\left(\sum_{i=1}^{n} X_i\right) - n\mu \geq n\epsilon\right) \leq \inf_{\lambda \geq 0} \frac{M_{\left(\sum_{i=1}^{n} X_i\right) - n\mu}(\lambda)}{exp(\lambda n\epsilon)}$

3. Simplify $M_{\left(\sum_{i=1}^{n} X_i\right) - n\mu}(\lambda)$ using Hoeffding's Lema:
   $M_{\left(\sum_{i=1}^{n} X_i\right) - n\mu}(\lambda) = M_{\sum_{i=1}^{n} X_i - \mu}(\lambda) = E[exp(\lambda(\sum_{i=1}^{n} X_i - \mu)] = E[\prod_{i=1}^{n} exp(\lambda(X_i - \mu)]$
   $\leq \prod_{i=1}^{n} exp\left(\frac{\lambda^2(b-a)^2}{8}\right) = exp\left(\frac{\lambda^2 n(b-a)^2}{8}\right)$

4. Subsituting back into the original equation, we get the following:
   $P\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu \geq \epsilon\right) = P\left(\left(\sum_{i=1}^{n} X_i\right) - n\mu \geq n\epsilon\right) \leq \inf_{\lambda \geq 0} \frac{M_{\left(\sum_{i=1}^{n} X_i\right) - n\mu}(\lambda)}{exp(\lambda n\epsilon)}$
   $\leq \inf_{\lambda \geq 0} \frac{exp\left(\frac{\lambda^2 n(b-a)^2}{8}\right)}{exp(\lambda n\epsilon)} = \inf_{\lambda \geq 0}\left(exp\left(\frac{\lambda^2 n(b-a)^2}{8} - \lambda n\epsilon\right)\right) = \inf_{\lambda \geq 0}\left(exp\left(\frac{\lambda^2 n(b-a)^2 - 8\lambda n\epsilon}{8}\right)\right)$

5. To find the optimatl λ take the derrivative and set it equal to zero
   $\frac{d}{d\lambda}\left(exp\left(\frac{\lambda^2 n(b-a)^2 - 8\lambda n\epsilon}{8}\right)\right) = \frac{2n(b-a)^2\lambda - 8n\epsilon}{8} exp\left(\frac{\lambda^2 n(b-a)^2 - 8\lambda n\epsilon}{8}\right) = 0$
   $\Rightarrow \frac{2n(b-a)^2\lambda}{8} exp\left(\frac{\lambda^2 n(b-a)^2 - 8\lambda n\epsilon}{8}\right) = (n\epsilon) exp\left(\frac{\lambda^2 n(b-a)^2 - 8\lambda n\epsilon}{8}\right)$
   $\Rightarrow \frac{2n(b-a)^2\lambda}{8} = n\epsilon$
   $\Rightarrow \lambda = \frac{8\epsilon}{2(b-a)^2} = \frac{4\epsilon}{(b-a)^2}$

6. Substituting this back into the original equation
   $\inf_{\lambda \geq 0}\left(exp\left(\frac{\lambda^2 n(b-a)^2}{8} - \lambda n\epsilon\right)\right) = exp\left(\frac{(\frac{4\epsilon}{(b-a)^2})^2 n(b-a)^2}{8} - (\frac{4\epsilon}{(b-a)^2})n\epsilon\right)$
   $= exp\left(\frac{2\epsilon^2 n}{(b-a)^2} - \frac{4\epsilon^2 n}{(b-a)^2}\right) = exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)$

7. This arrives at the final form of the proof:
   $P\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu \geq \epsilon\right) \leq exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)$

1.6 Applying Concentration inequalities to classifier evaluation

For any $1 > \delta > 0$, and $\epsilon > 0$, use Hoeffding's inequality to determine how many samples n we need to ensure that with probabiliyt at least 1 - δ, $|R_s(g) - R(g)|$ is less than ϵ

1. let $X_i = l_{01}(y, g(X)$ with range $[0, 1]$
2. Define $R(g) := E_D[l_{01}(y, g(x)) = E(X)$
3. Define $R_s(g) := \frac{1}{n}\sum_{i=1}^{n} l_{01}(y_i, g(x_i))$
4. Start by saying: $P(|R_s(g) - R(g)| < \epsilon) = 1 - P(|R_s(g) - R(g)| \geq \epsilon)$
5. Applying Hoeffding's inequality with $a = 0, b = 1$, we can then say:
   $1 - P(|R_s(g) - R(g)| \geq \epsilon) > 1 - 2exp(\frac{-2n\epsilon}{(b-a)^2}) = 1 - 2exp(-2n\epsilon)$

6. define $\delta := 2exp(-2n\epsilon)$
7. Simplifying further:
   $\delta = 2exp(-2n\epsilon^2)$
   $ln(\frac{\delta}{2}) = -2n\epsilon^2$
   $n = \frac{-1}{2\epsilon^2} ln(\frac{\delta}{2})$
   $n = \frac{1}{2\epsilon^2} ln(\frac{2}{\delta})$
8. Thus, in order to be a good approximation we n must meet the following condition
   $n > \frac{1}{2\epsilon^2} ln(\frac{2}{\delta})$

Next, we are asked to do the same thing with Chebychev's inequality

1. for $X = R_s(X)$, with each individual element having a variance of 1. We can use the properties of the variance to say that the variance of $Var(R_S(X)) = \sigma^2/n = \frac{1}{n}$
2. Next, we can say: $P(|R_s(g) - R(g)| < \epsilon) = 1 - P(|R_s(g) - R(g)| \geq \epsilon)$ as done before
3. Next, we can apply Chebyschev's inequality to say the following:
   $1 - P(|R_s(g) - R(g)| \geq \epsilon) > 1 - \frac{1}{n\epsilon^2}$
4. If we let $\delta = \frac{1}{n\epsilon^2}$, we can easily see that $n > \frac{1}{\delta\epsilon^2}$

Why is the sample complexity lower bound/guarantee given by Hoeffding's inequality prefered over the one given by Chebyschev's inequality
It is prefered because it decays faster versus the Chebyschev's Inequality

1.7 Application to Algorithmic Stability

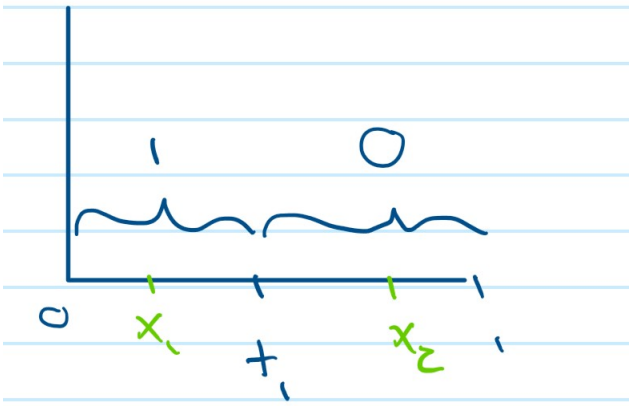
Ran out of time, did not complete

2 Classic Exercises in VC Dimmension

2.1

Find the VC dimension of the following hypothesis class. $\mathcal{F} = \{f : [0,1] \to \{0,1\}, f(x) = \mathbf{1}_{x<t}, t \in [0,1]\}$

The hypothesis class is plotted below

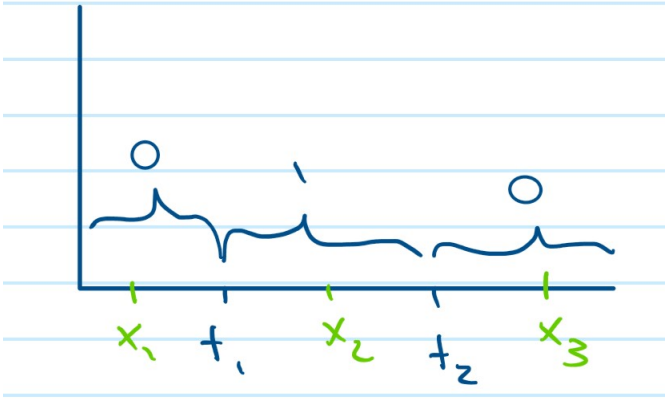


The VC dimension is the size of the largest set that a given classifier can shatter no matter what the labels are. As can be seen above, if $x_1 = 0, x_2 = 1$, the classifier cannot shatter the given set. As a result, the VC dimension is 1.

2.2

Find the VC dimension of the following hypothesis class. $\mathcal{F} = \{f : [0,1] \to \{0,1\}, f(x) = \mathbf{1}_{t_1<x<t_2}, t_1, t_2 \in [0,1]\}$

The hypothesis class is plotted below:



As can be seen above, if $x_1 = x_3 = 1, x_2 = 0$, the classifier will not be able to shatter the given set. As a result the VC dimension is 2.

2.3

Find the VC dimension of the following hypothesis class $\mathcal{F} = \{\mathbb{R}^d \to \{-1,1\}, f(\underline{x}) = sign(\underline{b}^T\underline{x}), \underline{b} \in \mathbb{R}^d\}$

1. Let $\mathcal{X} = \{\underline{e}_i\}_{i=1}^d$ where $\underline{e}_i$ are $d \times 1$ standard unit vectors.

(a) For a configuration of points as defined by $\mathcal{X}$, we can easily find a $\underline{b}$ that satisfies $f(\underline{x}) = sign(\underline{b}^T \underline{x})$. Let $\underline{b} = \sum_{i=1}^{d} y_i \underline{e}_i$.

(b) $\Rightarrow f(\underline{x}_i) = \sum_{i=1}^{d} y_i \underline{e}_i^T \underline{e}_i = y_i \underline{e}_i^T \underline{e}_i = y_i$

(c) Thus, we've shown that there is a classifier $\underline{b}$ that can shatter the configuration of d points described by $\mathcal{X}$ regardless of what the labels are.

2. However, if we were to add another point, $\underline{x}_{d+1} = \sum_{i=1}^{d} s_i \underline{x}_i$ for $x_i \in \mathcal{X}, s_i \in \mathbb{R}$, that is a linear combination of the elements in $\mathcal{X}$. However, we can show that there does not exist a $\underline{b}$ that can shatter the elements of the set regardless of label.

   (a) If, $\underline{b}$ can shatter the points of this new set, then the following would always hold: $y_{d+1} = f(x_{d+1}) = sign(\sum_{i=1}^{d} s_i) = sign(\underline{b}^T x_{d+1}) = sign(\underline{b}^T \sum_{i=1}^{d} s_i \underline{x}_i)$

   (b) However, this does not always hold. For example, if we say $s_i = \mathbf{1}_{y_i=1}$ and $y_i = -1$.

   (c) In this case $f(x_{d+1}) = sign(\underline{b}^T \sum_{i=1}^{d} s_i \underline{x}_i) > 0$.

   (d) However, since $y_i = -1$ which is a contradiction. Thus, we can say that there isn't a $\underline{b}$ that can shatter this slightly larger set.

3. The VC dimension is d

2.4

What is the VC dimension of the set of all binary decision trees with number of leaves at most l?

A decision tree can perfectly shatter at most l points if each point falls in a different leaf. However, if there are more points than leaves, then it may fail to perfectly classify each point (ex: two points with the same input, but different classifications). As a result, the VC dimension of all binary decision trees with l leaves is l.

What is the VC dimension of the set of all binary decision trees with number of splits at most d?

A decision tree with d splits will have $d + 1$ leaves. From the previous answer, this means that it will have a VC dimmension of $d + 1$.

2.5

Let $\mathcal{F}$ be a finite hypothesis class for binary classification, i.e $|\mathcal{F}| < \infty$. Show that the VC dimension of F is upper bounded by $log_2|\mathcal{F}|$

1. From the notes, the VC dimension of a class $\mathcal{F}$ is the largest n such that $\mathcal{S}_{\mathcal{F}}(n) = 2^n$ where $\mathcal{S}_{\mathcal{F}}(n)$ is the growth function .

2. Thus, for $|\mathcal{F}| < \infty$, the VC dimmension, will at most be the n such that $|\mathcal{F}| = 2^n$.

3. $\Rightarrow n = log_2|\mathcal{F}|$. Thus we have proved the upper bound.

3 Logistic Regression and Kernels

Let $\{(\underline{x}_i, y_i)\}_{i=1}^n$ be a set of training data where $\underline{x}_i \in \mathbb{R}^d$ for all i and $y_i \in \{-1, 1\}$.

Consider the l_2 regularized logistic regression model with parameter $\underline{\theta}$, where we want to find an $f_{\underline{\theta}}$ that minimizes the loss function:

$\sum_{i=1}^n ln(1 + exp(-y_i(f_{\underline{\theta}}(\underline{x}_i)))) + \lambda ||f_{\underline{\theta}}||_{\mathcal{H}}^2$

where $f_{\underline{\theta}}(\underline{x})$ is of the form $\underline{\theta}^T \underline{x}$ and $||f_{\underline{\theta}}||_{\mathcal{H}}^2 = \underline{\theta}^T \underline{\theta}$

3.1

Let $\mathcal{H}$ be the reproducing kernel Hilbert space that corresponds to the above l_2 regularized logistic regression. Using the representer theorem, what is the form of the optimal predictive function?

1. Per the notes on the representer theorem, for a fixed set $\mathcal{X}$, a kernel k and $\mathcal{H}$ as the corresponding RKHS. If $\Omega : \mathbb{R} \to \mathbb{R}$ is a nondecreasing function and any loss function $l : \mathbb{R}^d \to \mathbb{R}$, the solution to the optimation problem:
   $f^* \in argmin_{f \in \mathcal{H}} \sum_{i=1}^n l(f(x_i), y_i) + \Omega(||f||^2)$
   can be expressed in the form $f^* = \sum_{i=1}^n \alpha_i k(x_i, .)$
2. It can easily be seen that the optimization problem described in the problem statement easily aligns with the representer theorem.
3. As such, the form of the optimal predictive function will be $f^* = \sum_{i=1}^n \alpha_i k(x_i, .)$. In this case, based on the definition of $||f_{\underline{\theta}}||_{\mathcal{H}}^2 = \underline{\theta}^T \underline{\theta}$, we can say that the kernel is a linear kernel.

3.2 L2 Regularization

Define the following function g as $g(\zeta) = ln(1 + exp(-\zeta))$.
We are asked to solve the optimization problem specified by:
$\min_{\underline{w}, \underline{\zeta}} \frac{1}{2}||\underline{w}||^2 + C \sum_{i=1}^n g(\zeta_i)$
subject to:
$y_i(\underline{w}^T \underline{x}_i) \geq \zeta_i, \forall i$

1. Start by identifying the $g_i(x)$ constraints for the Lagrangian
   (a) $g_i(x)$ constraints come in the form $g_i(x) \leq 0$
   (b) We can put the given constraints into this form as follows:
   $y_i(\underline{w}^T \underline{x}_i) \geq \zeta_i$
   $0 \geq \zeta_i - y_i(\underline{w}^T \underline{x}_i)$
2. With the $g_i(x)$ constraints in the correct form, we can then form the lagrangian:
   $\mathcal{L}(\underline{w}, \underline{\zeta}, \underline{\alpha}) = \frac{1}{2}||\underline{w}||^2 + C \sum_{i=1}^n g(\zeta_i) + \sum_{i=1}^n \alpha_i [\zeta_i - y_i(\underline{w}^T \underline{x}_i)]$

3. Next we can apply the Lagrangian Stationarity Condition

   (a) With respect to the gradient of $\underline{w}$
   $\nabla_{\underline{w}}\mathcal{L}(\underline{w}, \underline{\zeta}, \underline{\alpha}) = \underline{w} - \sum_{i=1}^{n} \alpha_i y_i \underline{x}_i = 0$
   $\Rightarrow \underline{w} = \sum_{i=1}^{n} \alpha_i y_i \underline{x}_i$

   (b) With respect to the gradient of each ζ_i
   $\nabla_{\zeta_i}\mathcal{L}(\underline{w}, \underline{\zeta}, \underline{\alpha}) = \frac{-C}{1+exp(\zeta_i}$
   $\Rightarrow \alpha_i = \frac{C}{1+exp(\zeta_i)}$
   $\Rightarrow 1 + exp(\zeta_i) = \frac{C}{\alpha_i}$
   $\Rightarrow exp(\zeta_i) = \frac{C}{\alpha_i} - 1 = \frac{C-\alpha_i}{\alpha_i}$
   $\Rightarrow \zeta_i = ln(\frac{c-\alpha_i}{\alpha_i}) = -ln(\frac{\alpha_i}{C-\alpha_i})$
   $\Rightarrow 0 < \alpha_i < C$ where it is important to note that $\alpha_i > 0$

4. Apply the dual feasible condition which says that $\alpha_i \geq 0$. However, its also worth noting that we showed that $\alpha_i > 0$ when applying the Lagrangian Stationarity Condition

5. Apply Complementary Slackness Condition

   (a) $\alpha_i[\zeta_i - y_i(\underline{w}^T \underline{x}_i)] = 0$

   (b) Since, we showed that $\alpha_i > 0$ in the lagrangian stationarity condition, it follows that
   $\Rightarrow [\zeta_i - y_i(\underline{w}^T \underline{x}_i)] = 0$

6. Rewrite the Lagrangian and simplify with the conditions
   $\mathcal{L}(\underline{w}, \underline{\zeta}, \underline{\alpha}) = \frac{1}{2}||\underline{w}||^2 + C\sum_{i=1}^{n} g(\zeta_i) + \sum_{i=1}^{n} \alpha_i[\zeta_i - y_i(\underline{w}^T \underline{x}_i)]$

   (a) Simplify $\frac{1}{2}||\underline{w}||^2$ using the first lagrangian stationarity condition
   $\frac{1}{2}||\underline{w}||^2 = \frac{1}{2}\sum_{j=1}^{n}(w_i)^2 = \frac{1}{2}\sum_{j=1}^{n}(\sum_{i=1}^{n} \alpha_i y_i \underline{x}_{ij})^2 = \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\sum_{k=1}^{n} \alpha_i\alpha_k y_i y_k \underline{x}_{ij}\underline{x}_{kj}$
   $= \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n} \alpha_i\alpha_k y_i y_k \underline{x}_i^T \underline{x}_k$

   (b) Simplify $C\sum_{i=1}^{n} g(\zeta_i) = \sum_{i=1}^{n} ln(1+exp(-\zeta_i))$ using the second lagrangian stationary condition (with respect to ζ_i
   $\sum_{i=1}^{n} ln(1 + exp(-\zeta_i)) = \sum_{i=1}^{n} ln(1 + exp(ln(\frac{\alpha_i}{C-\alpha_i})))$
   $\sum_{i=1}^{n} ln(1 + \frac{\alpha_i}{C-\alpha_i})$

   (c) Finally, $\sum_{i=1}^{n} \alpha_i[\zeta_i - y_i(\underline{w}^T \underline{x}_i)]$ is always zero because the complementary slackness showed that $\alpha_i > 0$ which means that $[\zeta_i - y_i(\underline{w}^T \underline{x}_i)] = 0$.

7. Using the final simplifications, we arrive at the final form of the lagrangian for L2 regularization:
   $\mathcal{L}(\underline{\alpha}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n} \alpha_i\alpha_k y_i y_k \underline{x}_i^T \underline{x}_k + \sum_{i=1}^{n} ln(1 + \frac{\alpha_i}{C-\alpha_i}), 0 < \alpha_i < C$

4 Kernel Power on SVM and Regularized logistic Regression

5 Sparse Logistic Regression

6 Ridge Regression and its friends

6.1 Ridge Regression vs Kernel Ridge Regression

From my experiments, ridge regression took 2.1 seconds while kernel regression only took 0.6 seconds. The reason that ridge regression took so much longer than kernel regression can be seen in the closed form solutions for each solver.

For ridge regression, the closed form solution is $\lambda^* = (X^T X + CI)^{-1} X^T Y$. For an nxd X matrix where n is significantly smaller than d, the big-O complexity is $O(nd^2 + d^3)$

For kernel regression, the closed form solution is $\lambda^* = (XX^T + CI)^{-1} Y$. The big-O complexity is $O(dn^2 + n^3)$. Given that n is significantly larger than d, this helps to explain why kernel regression takes so less time than ridge regression in this example