Tutorials 4 & 5

CS 337 Artificial Intelligence & Machine Learning

Sunday 26th September, 2021

Problem 1. Weighted Linear Regression

Consider a data set in which each data point y_i is associated with a weighting factor r_i, so that the sum-square error function becomes

$$\frac{1}{2}\sum_{i=1}^{m} r_i(y_i - w^T\phi(x_i))^2$$

Find an expression for the solution w^* that minimizes this error function. The weights r_i's are known before hand. (Exercise 3.3 of Pattern Recognition and Machine Learning, Christopher Bishop).

Solution: Refer to the solution to problem 2, part 1. The solution to problem 1 is included therein.

Problem 2. Locally Weighted Kernel Regression

In problem 1, we discussed weighted regression. In this problem, we will deal with weighted regression, with the weights obtained using some kernel $K(.,.)$. Given a training set of points $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_n, y_n)\}$, we predict a regression function $f(x') = (\mathbf{w}^\top\phi(x') + b)$ for each test (or query point) x' as follows:

$$(\mathbf{w}', b') = \underset{\mathbf{w},b}{\operatorname{argmin}} \sum_{i=1}^{n} K(x', x_i)\left(y_i - (\mathbf{w}^\top\phi(x_i) + b)\right)^2$$

1. If there is a closed form expression for $(\mathbf{w}', b')$ and therefore for $f(x')$ in terms of the known quantities, derive it.
2. How does this model compare with linear regression and $k-$nearest neighbor regression? What are the relative advantages and disadvantages of this model?
3. In the one dimensional case (that is when $\phi(x) \in \Re$), graphically try and interpret what this regression model would look like, say when $K(.,.)$ is the linear kernel[1].

Solution:

This problem is directly related to problem 1 and herein we present the solution to both the problems

[1] Hint: What would the regression function look like at each training data point?

1. The weighing factor $r_i^{x'}$ of each training data point $(\mathbf{x}_i, y_i)$ is now also a function of the query or test data point $(\mathbf{x}', ?)$, so that we write it as $r_i^{x'} = K(\mathbf{x}', \mathbf{x}_i)$ for $i = 1, \ldots, m$. Let $r_{m+1}^{x'} = 1$ and let R be an $(m+1) \times (m+1)$ diagonal matrix of $r_1^{x'}, r_2^{x'}, \ldots, r_{m+1}^{x'}$.

$$R = \begin{bmatrix} r_1^{x'} & 0 & ... & 0 & \\ 0 & r_2^{x'} & ... & 0 & \\ ... & ... & ... & ... & 1 \\ 0 & 0 & 0 & ... & r_{m+1}^{x'} \end{bmatrix}$$

Further, let

$$\Phi = \begin{bmatrix} \phi_1(x_1) & ... & \phi_p(x_1) & 1 \\ ... & ... & ... & 1 \\ \phi_1(x_m) & ... & \phi_p(x_m) & 1 \end{bmatrix}$$

and

$$\widehat{\mathbf{w}} = \begin{bmatrix} w_1 \\ ... \\ w_p \\ b \end{bmatrix}$$

and

$$\mathbf{y} = \begin{bmatrix} y_1 \\ ... \\ y_m \end{bmatrix}$$

The sum-square error function then becomes

$$\frac{1}{2}\sum_{i=1}^{m} r_i(y_i - (\widehat{\mathbf{w}}^T\phi(x_i) + b))^2 = \frac{1}{2}||\sqrt{R}\mathbf{y} - \sqrt{R}\Phi\widehat{\mathbf{w}}||_2^2$$

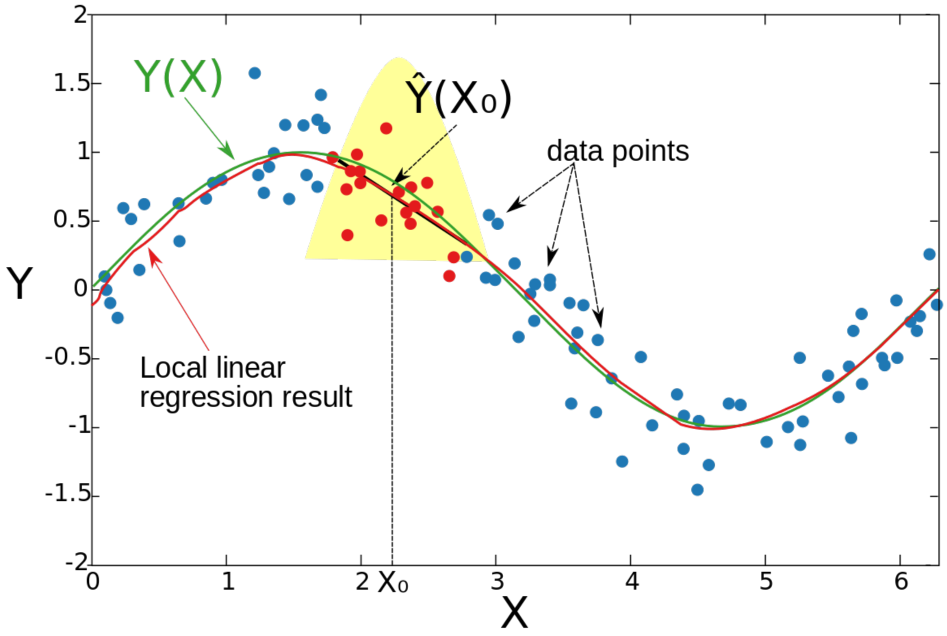
where $\sqrt{R}$ is a diagonal matrix such that each diagonal element of $\sqrt{R}$ is the square root of the corresponding element of R. This is a convex function being minimized (prove this using techniques similar to what we employed for least squares linear regression) and therefore has a global minimum at $\widehat{\mathbf{w}}_*^{x'}$ where the gradient must become 0. (again work out the steps using techniques similar to what we employed for least squares linear regression). The expression for the solution $\widehat{\mathbf{w}}^*$ that minimizes this error function is therefore

$$\widehat{\mathbf{w}}_*^{x'} = (\Phi^T R \Phi)^{-1}\Phi^T R\mathbf{y}$$

2. Let us refer to this model as local linear regression (Section 6.1.1 of Tibshi's book).

   As compared to linear regression, local linear regression gives more importance to points in $\mathcal{D}$ that are closer/similar to $\mathbf{x}'$ and less importance to points that are less similar. Thus, this method can be important if the regression curve is supposed to take different shapes or different parameters in different parts of the space. For example, in two different regions, the ideal regression curve might be linear in each but with different parameters. In this sense, local linear regression comes close to k-nearest neighbor. But unlike k-nearest neighbor, local linear regression gives you a smooth solution since contribution for regression at a point comes from all data points (in proportion to their closeness) and not just the k closest points.

3. Taking clue from the discussion above, one can try and plot this regression curve.



Problem 3. Redoing the Kernel Ridge Regression Problem: Let $\mathcal{D} = \langle(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\rangle$ such that each $y_j \in \Re$. Let $\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_n(\mathbf{x})]$ be a vector of basis functions. Consider the linear regression function $f(\mathbf{x}) = \phi^T(\mathbf{x})\mathbf{w}$ with $\mathbf{w}$ obtained either as a least squares or ridge regression estimate. Show that, using either of these estimates for $\mathbf{w}$, the regression function can be written in the (so-called *kernelized*) form $f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}_i) y_i$ where $K(\mathbf{x}, \mathbf{x}_i) = \phi^T(\mathbf{x})\phi(\mathbf{x}_i)$ is a function of $\mathbf{x}$ and $\mathbf{x}_i$ only and independent of any of the $\mathbf{y}_i$'s and $\mathbf{x}_j$ for all $j \neq i$. Each α_i can be a function of the entire dataset $\mathcal{D}$.

Hint: Use the following Matrix Identity that holds for any matrices P, B and R with compatible dimensions such that R and $BPB^T + R$ are invertible:

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = PB^T (BPB^T + R)^{-1}$$

Solution: The solution to linear (set $\lambda = 0$) and ridge regression can be written as $\mathbf{w} = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T \mathbf{y}$ where

- Recall for Ridge Regression: $\mathbf{w} = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T y$, where,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \ldots \\ y_m \end{bmatrix}$$

$$\Phi = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \ldots & \phi_p(\mathbf{x}_1) \\ \ldots & \ldots & \ldots \\ \phi_1(\mathbf{x}_m) & \ldots & \phi_p(\mathbf{x}_m) \end{bmatrix}$$

- **Please note the difference between Φ and $\phi(\mathbf{x})$**

$$\phi(\mathbf{x}_j) = \begin{bmatrix} \phi_1(\mathbf{x}_j) \\ ... \\ \phi_p(\mathbf{x}_j) \end{bmatrix}$$

  Then, the regression function will be

$$f(\mathbf{x}) = \phi^T(\mathbf{x})\mathbf{w} = \phi^T(\mathbf{x})(\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y}$$

- $\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$
- $\left(\Phi^T\Phi\right)_{ij} = \sum_{k=1}^m \phi_i(\mathbf{x}_k)\phi_j(\mathbf{x}_k)$
- $\left(\Phi\Phi^T\right)_{ij} = \sum_{k=1}^p \phi_k(\mathbf{x}_i)\phi_k(\mathbf{x}_j) = \phi^T(x_i)\phi(x_j) = K(\mathbf{x}_i, \mathbf{x}_j)$

Kernelizing Ridge Regression

- Given $\mathbf{w} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y}$ and using the identity $(P^{-1} + B^TR^{-1}B)^{-1}B^TR^{-1} = PB^T(BPB^T + R)^{-1}$
  - $\Rightarrow$ by setting $R = I$, $P = \frac{1}{\lambda}I$ and $B = \Phi$,
  - $\Rightarrow$ $\mathbf{w} = \Phi^T(\Phi\Phi^T + \lambda I)^{-1}\mathbf{y} = \sum_{i=1}^m \alpha_i\phi(\mathbf{x}_i)$ where $\alpha_i = \left((\Phi\Phi^T + \lambda I)^{-1}\mathbf{y}\right)_i$
  - $\Rightarrow$ the final decision function $f(\mathbf{x}) = \phi^T(\mathbf{x})\mathbf{w} = \sum_{i=1}^m \alpha_i\phi^T(\mathbf{x})\phi(\mathbf{x}_i)$

The Kernel function in Ridge Regression

- We call $\phi^\top(x_1)\phi(x_2)$ a **kernel function**:
  $K(x_1, x_2) = \phi^\top(x_1)\phi(x_2)$
- The preceding expression for decision function becomes $f(\mathbf{x}) = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i)$ where $\alpha_i = (([K(\mathbf{x}_i, \mathbf{x}_j)] + \lambda I)^{-1}\mathbf{y})_i$

Problem 4. Equivalent Kernelized Representation (Post-midsem):

Throughout this question, let $0 < p < 1$. Consider a data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ of m points and a feature function $\phi(\mathbf{x}) \subseteq \Re^n$. Let $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$. You have seen linear regression with various regularizations of the form:

$$\sum_{i=1}^m \left(y_i - \phi^T(\mathbf{x}_i)\mathbf{w}\right)^2 + \lambda\left(\sum_{j=1}^n |w_j|^p\right) \tag{1}$$

Now consider a somewhat complementary setting:

$$\sum_{i=1}^m \left(y_i - \phi^T(\mathbf{x}_i)\mathbf{w}\right)^p + \lambda\left(\sum_{j=1}^n |w_j|^2\right) \tag{2}$$

1. Do these forms have an equivalent kernelized representation: $f(\mathbf{x}) = \phi^T(\mathbf{x})\mathbf{w}^* + b = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$? How would you prove?

2. Contrast the two descriptions for their capabilities.

Solution:
Solution to part (1):

As per the representer theorem, if $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$ and $K(\mathbf{x}', \mathbf{x}) = \phi^T(\mathbf{x})\phi(\mathbf{x}')$ then the solution $\mathbf{w}^* \in \Re^n$ to the following problem

$$(\mathbf{w}^*, b^*) = \underset{\mathbf{w},b}{\operatorname{argmin}} \sum_{i=1}^{m} \mathbf{E}\left(f\left(\mathbf{x}^{(i)}\right), y^{(i)}\right) + \Omega(||\mathbf{w}||_2)$$

can be always written as $\phi^T(\mathbf{x})\mathbf{w}^* + b = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$, provided $\Omega(||\mathbf{w}||_2)$ is a monotonically increasing function of $||\mathbf{w}||_2$. **Optionally recall:** $\Re^n$ is the Hilbert space and $K(., \mathbf{x}) : \mathcal{X} \to \Re$ is the **Reproducing (RKHS) Kernel**

(it is ok if the student does not mention Hilbert space etc).

Thus, it is obvious that the errors of both (5) and (6) linearly decompose across all the examples as expected in the Representer theorem, the regularizer in (5) cannot be written as a monotonically increasing function of $||\mathbf{w}||_2$ which is possible in the case of (6). Thus, only (6) has an equivalent kernelized representation whereas (5) does not have one.

Solution to part (2):

Whereas (5) gives relatively sparser $\mathbf{w}$ than (6), (6) takes care of outliers better than (5) by having a slower rate of growth with respect to the error (recall this discussion from Tutorial 1).

Problem 5. More on Kernel Perceptron:

Recall the proof for convergence of the perceptron update algorithm. Now can this proof be extended to the kernel perceptron?

Recall that Kernelized perceptron is specified as:

$f(x) = sign\left(\sum_i \alpha_i^* y_i K(x, x_i)\right)$

The perceptron update algorithm for the Kernelized version is:

- INITIALIZE: α=zeroes()
- REPEAT: for $< x_i, y_i >$
  - If $sign\left(\sum_j \alpha_j y_j K(x_j, x_j)\right) \neq y_i$
  - then, $\alpha_j = \alpha_j + 1$

Solution: Yes, in fact kernel perceptron can be derived from the perceptron update rule as follows:

$f(x) = sign\left((w^*)^T\phi(x)\right) = sign\left(\sum_i \alpha_i^* y_i K(x, x_i)\right)$

- INITIALIZE: $w = [0, 0, \ldots, 0, 1] \Rightarrow f(x) = sign\left((w)^T\phi(x)\right) = sign\left(\sum_i \alpha_i y_i K(x, x_i)\right)$

  with $\alpha_i = 0$

  Note: $\phi^T(\widehat{x})\phi(x)\widehat{y} = \widehat{y}K(\widehat{x}, x) + \widehat{y}$

- REPEAT: for each $<\widehat{x}, \widehat{y}>$

  - If $\widehat{y}w^T\phi(\widehat{x}) < 0$

    $\Rightarrow f(\widehat{x}) = sign\left((w)^T\phi(\widehat{x})\right) = sign\left(\sum_i \alpha_i y_i K(\widehat{x}, x_i)\right) \neq \widehat{y}$

  - then, $w' = w + \Phi(\widehat{x}).\widehat{y}$

    $\Rightarrow f(x) = sign\left((w')^T\phi(x)\right) = sign\left(\sum_i (\alpha_i y_i K(x, x_i) + \phi^T(\widehat{x})\phi(x)\widehat{y})\right) = sign\Bigg(\sum_i \alpha'_i y_i K(x, x_i$

    where $\alpha'_i = \alpha_i$ for all i except that $\alpha'_{\widehat{x}} = \alpha_{\widehat{x}} + 1$.

  - endif

Thus, $f(x) = sign\left((w^*)^T\phi(x)\right) = sign\left(\sum_i^* \alpha_i y_i K(x, x_i)\right)$ Having proved this mapping from every step of kernel perceptron update rule to the regular perceptron update rule, it suffices to say that the proof of convergence for the perceptron update rule will therefore hold even for the kernel perceptron update rule.

Problem 6. Kernel Logistic Regression: Intuition (and optional Rigorous proof)

Recall the Regularized (Logistic) Cross-Entropy Loss function (minimized wrt $\mathbf{w} \in \Re^p$):

$$E(\mathbf{w}) = -\left[\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\log f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right) + \left(1 - y^{(i)}\right)\log\left(1 - f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right)\right)\right)\right] + \frac{\lambda}{2m}\|\mathbf{w}\|_2^2 \qquad (3)$$

Now intuitively show that minimizing the following dual kernelized objective[2] (minimized wrt $\alpha \in \Re^m$) is equivalent to minimizing the regularized cross-entropy loss function:

$$E_D(\alpha) = \left[\sum_{i=1}^{m}\left(\sum_{j=1}^{m} -y^{(i)}K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\alpha_j + \frac{\lambda}{2}\alpha_i K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\alpha_j\right) + \log\left(1 + \exp\sum_{j=1}^{m}\alpha_j K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\right)\right] \qquad (4)$$

where, decision function $f_{\mathbf{w}}(\mathbf{x}) = \dfrac{1}{1+\exp\left(-\sum_{\mathbf{j=1}}^{\mathbf{m}}\alpha_{\mathbf{j}}\mathbf{K}\left(\mathbf{x}, \mathbf{x}^{(\mathbf{j})}\right)\right)}$ How would you prove this very rigorously (**optional**)?

Solution:

We will prove this result and in the process, also motivate (and somewhat prove - **optional**) the more general **Representer Theorem** atleast for Logistic Regression

[2] http://perso.telecom-paristech.fr/~clemenco/Projets_ENPC_files/kernel-log-regression-svm-boosting.pdf

1. **Some preliminary steps:**

   Recall another form of the regularized cross entropy equivalent to (3)

$$E\left(\mathbf{w}\right) = -\left[\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\mathbf{w}^{T}\phi(\mathbf{x}^{(i)}) - \log\left(1+\exp\left(\mathbf{w}^{T}\phi(\mathbf{x}^{(i)})\right)\right)\right)\right] + \frac{\lambda}{2m}||\mathbf{w}||^{2} \tag{5}$$

   First of all, we will drop the common term $\frac{1}{m}$ from the primal optimization problem and equivalently minimize the unscaled version

$$E\left(w\right) = -\left[\sum_{i=1}^{m}\left(y^{(i)}\log f_{w}\left(\mathbf{x}^{(i)}\right) + \left(1-y^{(i)}\right)\log\left(1-f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right)\right)\right)\right] + \frac{\lambda}{2}\|\mathbf{w}\|_{2}^{2} \tag{6}$$

$$\nabla E\left(\mathbf{w}\right) = \left[\sum_{i=1}^{m}\left(y^{(i)}\nabla\log f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right) + \left(1-y^{(i)}\right)\nabla\log\left(1-f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right)\right)\right)\right] + \lambda\mathbf{w} \tag{7}$$

2. $\nabla\log f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right) = \phi(\mathbf{x}^{(i)})e^{-(\mathbf{w})^{T}\phi(\mathbf{x}^{(i)})}\left(\frac{1}{1+e^{-(\mathbf{w})^{T}\phi(\mathbf{x}^{(i)})}}\right)^{2}$ and

   $\nabla\log\left(1-f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right)\right) = -\phi(\mathbf{x}^{(i)})\left(\frac{1}{1+e^{-(\mathbf{w})^{T}\phi(\mathbf{x}^{(i)})}}\right)^{2}$

3. ⇒

$$\nabla E\left(\mathbf{w}\right) = \left[\sum_{i=1}^{m}\left(y^{(i)} - f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right)\right)\phi(\mathbf{x}^{(i)})\right] + \lambda\mathbf{w} \tag{8}$$

   At optimality, a necessary condition is that $\nabla E\left(\mathbf{w}\right) = 0$ and therefore,

$$\mathbf{w} = \frac{1}{\lambda}\left[\sum_{i=1}^{m}\left(y^{(i)} - f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right)\right)\phi(\mathbf{x}^{(i)})\right] \tag{9}$$

4. **The main idea:**

   In summary, the main optimization objective is

$$E\left(\mathbf{w}\right) = -\left[\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\mathbf{w}^{T}\phi(\mathbf{x}^{(i)}) - \log\left(1+\exp\left(\mathbf{w}^{T}\phi(\mathbf{x}^{(i)})\right)\right)\right)\right] + \frac{\lambda}{2m}||\mathbf{w}||^{2} \tag{10}$$

   and an expression for **w** at optimality is

$$\mathbf{w} = \frac{1}{\lambda}\left[\sum_{i=1}^{m}\left(y^{(i)} - f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right)\right)\phi(\mathbf{x}^{(i)})\right] \tag{11}$$

5. Recall from the representer theorem that in the optimization problem (10), $\mathbf{w^T}\phi(\mathbf{x^{(i)}})$ can be equivalently expressed as $\sum_{\mathbf{j=1}}^{\mathbf{m}} \alpha_{\mathbf{j}}\mathbf{K}\left(\mathbf{x}, \mathbf{x^{(j)}}\right)$, as a result of which we will obtain the following terms of (4):

$$\left[\sum_{i=1}^{m}\left(\sum_{j=1}^{m} -y^{(i)}K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\alpha_j\right) + \log\left(1 + \exp\sum_{j=1}^{m}\alpha_j K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\right)\right] \tag{12}$$

Substituting from (11) into one of the $\mathbf{w}$ in $\frac{\lambda}{2m}||\mathbf{w}||^2 = \frac{\lambda}{2m}\mathbf{w}^T\mathbf{w}$ term of (10) we will get the regularizer into the form

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\lambda}{2}\alpha_i K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\alpha_j$$

which helps form the remaining term of (4)

EXTRA and COMPLETELY OPTIONAL: Proof of Representer Theorem specifically for Logistic Regression:

To completely prove this specific case of KLR, let $\mathcal{X}$ be the space of examples such that $\left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}\right\} \subseteq \mathcal{X}$ and for any $\mathbf{x} \in \mathcal{X}$, $K(.,\mathbf{x}) : \mathcal{X} \to \Re$ be a function such that $K(\mathbf{x}', \mathbf{x}) = \phi^T(\mathbf{x})\phi(\mathbf{x}')$. Recall that $\phi(\mathbf{x}) \in \Re^n$ and

$$f_{\mathbf{w}}(\mathbf{x}) = p(Y = 1|\phi(\mathbf{x})) = \frac{1}{1 + \exp\left(-\mathbf{w^T}\phi(\mathbf{x})\right)}$$

For the rest of the discussion, we are interested in viewing $-\mathbf{w}^T\phi(\mathbf{x})$ as a function $h(\mathbf{x})$

$$f_{\mathbf{w}}(\mathbf{x}) = p(Y = 1|\phi(\mathbf{x})) = \frac{1}{1 + \exp\left(\mathbf{h}(\mathbf{x})\right)}$$

We will prove that for the optimization problem (10), $\mathbf{h}(\mathbf{x})$ can be equivalently expressed as $\sum_{\mathbf{j=1}}^{\mathbf{m}} \alpha_{\mathbf{j}}\mathbf{K}\left(\mathbf{x}, \mathbf{x^{(j)}}\right)$, as a result of which we will obtain the following terms of (4):

$$\left[\sum_{i=1}^{m}\left(\sum_{j=1}^{m} -y^{(i)}K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\alpha_j\right) + \log\left(1 + \sum_{j=1}^{m}\alpha_j K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\right)\right] \tag{13}$$

Substituting (11) into $\frac{\lambda}{2m}||\mathbf{w}||^2$ term of (10) we will get the regularizer into the form

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\lambda}{2}\alpha_i K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\alpha_j$$

which forms the remaining term of (4)

1. Consider the set of functions $\mathcal{K} = \{K(.,\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$ and let $\mathcal{H}$ be the set of all functions that are **finite** linear combinations of functions in $\mathcal{K}$. That is, any function $h \in \mathcal{H}$ can

be written as $\mathbf{h}(.) = \sum_{t=1}^{T} \alpha_t K(., \mathbf{x}_t)$ for some T and $\mathbf{x}_t \in \mathcal{X}, \alpha_t \in \Re$. One can easily verify that $\mathcal{H}$ is a vector space[3]

Note that, in the special case when $f(\mathbf{x}') = K(\mathbf{x}', \mathbf{x})$, then $T = m$ and

$$f(\mathbf{x}') = K(\mathbf{x}', \mathbf{x}) = \sum_{i=1}^{n} \phi_i(\mathbf{x}') K(\mathbf{e}_i, \mathbf{x})$$

where $\mathbf{e}_i$ is such that $\phi(\mathbf{e}_i) = \mathbf{u}_i \in \Re^n$, the unit vector along the i^{th} direction.

Also, by the same token, if $\mathbf{w} \in \Re^n$ is in the search space of the regularized cross-entropy loss function (3), then

$$\phi^{\mathbf{T}}(\mathbf{x}')\mathbf{w} = \sum_{i=1}^{n} w_i K(\mathbf{e}_i, \mathbf{x})$$

Thus, the solution to (3) is an $h \in \mathcal{H}$.

2. **Inner Product over $\mathcal{H}$:** For any $g(.) = \sum_{t=1}^{S} \beta_s K(., \mathbf{x}'_s) \in \mathcal{H}$ and $h(.) = \sum_{t=1}^{T} \alpha_t K(., \mathbf{x}_t) \in \mathcal{H}$, define the inner product[4]

$$\langle h, g \rangle = \sum_{s=1}^{S} \beta_s \sum_{t=1}^{T} \alpha_t K(\mathbf{x}'_s, \mathbf{x}_t) \tag{14}$$

Further simplifying (14),

$$\langle h, g \rangle = \sum_{s=1}^{S} \beta_s \sum_{t=1}^{T} \alpha_t K(\mathbf{x}'_s, \mathbf{x}_t) = \sum_{s=1}^{S} \beta_s f(\mathbf{x}_s) \tag{15}$$

One immediately observes that in the special case that $g() = K(., \mathbf{x})$,

$$\langle h, K(., \mathbf{x}) \rangle = h(\mathbf{x}) \tag{16}$$

3. **Orthogonal Decomposition:** Since $\left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}\right\} \subseteq \mathcal{X}$ and $\mathcal{K} = \{K(., \mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$ with $\mathcal{H}$ being the set of all finite linear combinations of function in $\mathcal{K}$, we also have that

$$lin\_span \left\{K(., \mathbf{x}^{(1)}), K(.\mathbf{x}^{(2)}), \ldots, K(., \mathbf{x}^{(m)})\right\} \subseteq \mathcal{H}$$

[3] Try it yourself. Prove that $\mathcal{H}$ is closed under vector addition and (real) scalar multiplication.

[4] Again, you can verify that $\langle f, g \rangle$ is indeed an inner product following properties such as symmetry, linearity in the first argument and positive-definiteness: https://en.wikipedia.org/wiki/Inner_product_space

Thus, we can use orthogonal projection to decompose any $h \in \mathcal{H}$ into a sum of two functions, one lying in $lin\_span\left\{K(.,\mathbf{x}^{(1)}), K(.\mathbf{x}^{(2)}), \ldots, K(.,\mathbf{x}^{(m)})\right\}$, and the other lying in the orthogonal complement:

$$h = h^{\parallel} + h^{\perp} = \sum_{i=1}^{m} \alpha_i K(.,\mathbf{x}^{(i)}) + h^{\perp} \tag{17}$$

where $\langle K(.,\mathbf{x}^{(i)}), h^{\perp}\rangle = 0$, for each $i = [1..m]$.

For a specific training point $\mathbf{x}^{(j)}$, substituting from (17) into (16) for any $h \in \mathcal{H}$, using the fact that $\langle K(.,\mathbf{x}^{(i)}), h^{\perp}\rangle = 0$

$$h(\mathbf{x}^{(j)}) = \langle \sum_{i=1}^{m} \alpha_i K(.,\mathbf{x}^{(i)}) + h^{\perp}, K(.,\mathbf{x}^{(j)})\rangle = \sum_{i=1}^{m} \alpha_i \langle K(.,\mathbf{x}^{(i)}), K(.,\mathbf{x}^{(j)})\rangle = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \tag{18}$$

which we observe is independent of $h^{\perp}$.

4. **Analysis of the Regularized Cross-Entropy Logistic Loss:**

   The Regularized Cross-Entropy Logistic Loss (10), has two parts (after ignoring the common $\frac{1}{m}$ factor), *viz.*, the **empirical risk**

$$-\left[\sum_{i=1}^{m} \left(y^{(i)}\mathbf{w}^T\phi(\mathbf{x}^{(i)}) - \log\left(1 + \exp\left(\mathbf{w}^T\mathbf{x}^{(i)}\right)\right)\right)\right] \tag{19}$$

   Since the **empirical risk** in (19) is only a function of $h(\mathbf{x}^{(i)}) = \mathbf{w}^T\phi(\mathbf{x}^{(i)})$ for $i = [1..m]$, based on (18) we note that the value of the **empirical risk** in (19) will therefore be independent of $h^{\perp}$ and therefore one only needs to equivalently solve the following **empirical risk** by substituting from (18) *i.e.*, $h(\mathbf{x}^{(j)}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$:

$$\left[\sum_{\mathbf{i=1}}^{\mathbf{m}}\left(\sum_{\mathbf{j=1}}^{\mathbf{m}} -\mathbf{y^{(i)}K}\left(\mathbf{x^{(i)}}, \mathbf{x^{(j)}}\right)\alpha_{\mathbf{j}}\right) + \log\left(\mathbf{1} + \sum_{\mathbf{j=1}}^{\mathbf{m}} \alpha_{\mathbf{j}}\mathbf{K}\left(\mathbf{x^{(i)}}, \mathbf{x^{(j)}}\right)\right)\right]$$

5. **Safe with Regularizer?**

   Consider the regularizer function $||\mathbf{w}||_2^2$ which is a strictly monotonically increasing function of $||\mathbf{w}||$. Substituting $\mathbf{w} = \frac{1}{\lambda}\left[\sum_{i=1}^{m}\left(y^{(i)} - f_{\mathbf{w}}\left(\mathbf{x}^{(i)}\right)\right)\phi(\mathbf{x}^{(i)})\right]$ from (9), one can view $\Omega(||h||)$ as a strictly monotonic function of $||h||$.

$$\Omega(||h||) = \Omega\left(||\sum_{i=1}^{m} \alpha_i K(.,\mathbf{x}^{(i)}) + h^{\perp}||\right) = \Omega\left(\sqrt{||\sum_{i=1}^{m} \alpha_i K(.,\mathbf{x}^{(i)})||^2 + ||h^{\perp}||^2}\right)$$

and therefore,

$$\Omega(||h||) = \Omega\left(\sqrt{||\sum_{i=1}^{m}\alpha_i K(.,\mathbf{x}^{(i)})||^2 + ||h^{\perp}||^2}\right) \geq \Omega\left(\sqrt{||\sum_{i=1}^{m}\alpha_i K(.,\mathbf{x}^{(i)})||^2}\right)$$

That is, setting $h^{\perp} = 0$ does not affect the first term of (10) while strictly increasing the second term. That is, any minimizer must have optimal $h^*(.)$ with $h^{\perp} = 0$. That is,

$$h(\mathbf{x}) = \sum_{i=1}^{m}\alpha_i K(\mathbf{x}^{(i)},\mathbf{x})$$

Problem 7. Effect of increasing λ in Ridge Regression

Consider the ridge regression problem

$$\widehat{\mathbf{w}}_{ridge} = \underset{\mathbf{w}}{\text{argmin}}\ ||\Phi^T\mathbf{w} - \mathbf{y}||^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

for any $\lambda \geq 0$. Recall the purpose for which the regularization term $\frac{\lambda}{2}||\mathbf{w}||^2$ was introduced in linear regression.

Let $\mathbf{w}_1$ be the optimal solution to this problem when $\lambda = \lambda_1$ and let $\mathbf{w}_2$ be the optimal solution to this problem when $\lambda = \lambda_2$. Let $\lambda_2 < \lambda_1$.

Which of the following statements is correct?

1. $||\mathbf{w}_2|| \leq ||\mathbf{w}_1||$ (that is, $||\widehat{\mathbf{w}}_{ridge}||$ will not increase as λ decreases towards 0).
2. $||\mathbf{w}_2|| \geq ||\mathbf{w}_1||$ (that is, $||\widehat{\mathbf{w}}_{ridge}||$ will not decrease as λ decreases towards 0).
3. none of these

Prove your answer. Why does increase in λ reduce the curvature of the solution obtained via ridge regression?

SOLUTION:

Let $\lambda_2 < \lambda_1$ and $f_\lambda(\mathbf{w}) = ||\Phi^T\mathbf{w} - \mathbf{y}||^2 + \frac{\lambda}{2}||\mathbf{w}||^2$ Then:

$$f_{\lambda_1}(\mathbf{w}_2) + f_{\lambda_2}(\mathbf{w}_1) \geq f_{\lambda_1}(\mathbf{w}_1) + f_{\lambda_2}(\mathbf{w}_2)$$

i.e,

$$\lambda_1||\mathbf{w}_2||^2 + \lambda_2||\mathbf{w}_1||^2 \geq \lambda_1||\mathbf{w}_1||^2 + \lambda_2||\mathbf{w}_2||^2$$

i.e

$$(\lambda_1 - \lambda_2)||\mathbf{w}_2||^2 \geq (\lambda_1 - \lambda_2)||\mathbf{w}_1||^2$$

and since $\lambda_2 < \lambda_1$

$$||\mathbf{w}_2||^2 \geq ||\mathbf{w}_1||^2$$

That is, $||\mathbf{w}_2||$ increases as λ decreases

Look at the countours of the objective $||A\mathbf{x}-\mathbf{b}||^2$. The larger is the ratio $\frac{\lambda_{max}(A)}{\lambda_{min}(A)}$, the more skewed are the level curves and more is the time gradient descent will take for convergence. Thus, the matrix A with small value of $\frac{\lambda_{max}(A)}{\lambda_{min}(A)}$ is always desirable.

In general, by the Courant-Fischer min-max Theorem, if A and B are two $n\times n$ symmetric matrices, and suppose the k^{th} largest eigenvalue of matrix X is $\lambda_k(X)$, $k = 1, 2, \ldots, n$: $\lambda_1(X) \geq \lambda_2(X) \ldots \geq \lambda_n(X)$ then

$\lambda_k(A) + \lambda_n(B) \leq \lambda_k(A+B) \leq \lambda_k(A) + \lambda_1(B)$

Problem 8. Are these Valid Kernels? Consider the space of all possible subsets A of a given fixed set D. Prove/disprove the following functions are valid Kernels:

1. $K(A_1, A_2) = |A_1 \cap A_2|$
2. $K(A_1, A_2) = 2^{|A_1 \cap A_2|}$

where A_1, A_2 are subsets of D and $|B|$ is the cardinality of $|B|$ or the number of elements in B.

Solution:

Solution to part (a):

Construct a feature vector of size equal to the cardinality of A with $0/1$ for an element of set being absent/present respectively in a subset. Taking the product of two feature vectors would give the number of common elements because only the 1's corresponding to elements present in both subsets give a multiplicative contribution.

Solution to part (b):

One way is to prove that if $K(A_1, A_2)$ is a Kernel then $2^{K(A_1,A_2)}$ is also a Kernel. This can be done using the power series expansion of 2^x where coefficient of each degree is positive and then invoking that positive polynomials of Kernels are also Kernels.

Other way is to explicitly come up with a feature vector representation for $2^{|A_1 \cap A_2|}$. Consider feature vector of size $2^{|A|}$ defined by mapping $\phi(A_1)$ where A_1 is a subset of A and each component is representative of a possible subset U of A with entry either 0 or 1 for some subset A_1 as follows:

$$\phi_U(A_1) = 1, \text{if } U \subseteq A_1$$
$$\phi_U(A_1) = 0, \text{otherwise}$$