# Tutorials 4 & 5

## CS 337 Artificial Intelligence & Machine Learning

Sunday 26$^{\text{th}}$ 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).

**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:

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

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].

**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

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

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 $\alpha_i$ 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:

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

**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.

**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: $\alpha$=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, $\alpha_j = \alpha_j + 1$

**Problem 6. Kernel Logistic Regression: Intuition (and optional Rigorous proof)** Recall the Regularized (Logistic) Cross-Entropy Loss function (minimized wrt $\mathbf{w} \in \Re^p$):

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

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:

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

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

**Problem 7. Effect of increasing $\lambda$ in Ridge Regression**

Consider the ridge regression problem

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

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

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

**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 $B$.

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