Engineering Applications of Machine Learning and Data Analytics
Homework #2

I acknowledge that this exam is solely my effort. I have done this work by myself. I have not consulted with others about this exam in any way. I have not received outside aid (outside of my own brain) on this exam. I understand that violation of these rules contradicts the class policy on academic integrity.

Name: _____

Signature: _____

Date: _____

Instructions: There are four problems. X Partial credit is given for answers that are partially correct. No credit is given for answers that are wrong or illegible. Write neatly.

You must submit two PDFs on D2L. The first PDF has the results to the analytical questions as well as figures that are generated

Problem 1: _____ Problem 2: _____

Problem 3: _____ Problem 4: _____

Total: _____

1 Linear Classifier with a Margin [10pts]

Show that, regardless of the dimensionality of the feature vectors, a data set that has just two data points, one from each class, is sufficient to determine the location of the maximum-margin hyperplane. Hint #1: Consider a data set of two data points, $\mathbf{x}_1 \in \mathcal{C}_1$ ($y_1 = +1$) and $\mathbf{x}_2 \in \mathcal{C}_2$ ($y_2 = -1$) and set up the minimization problem (for computing the hyperplane) with appropriate constraints on $\mathbf{w}^\mathsf{T}\mathbf{x}_1 + b$ and $\mathbf{w}^\mathsf{T}\mathbf{x}_2 + b$ and solve it. Hint #2: This can be formed as a constrained optimization problem.

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \|\mathbf{w}\|_2^2$$

Subject to: (some constraint)

What is $\mathbf{w}$? b? Hint: What are the constraints? How did we solve the constrained optimization problem in Fisher's linear discriminate (see Linear Models Lecture Notes or constrained optimization from Calculus)?

Solution Irrespective of the dimensionality of the data space, a data set consisting of just two data points – one from each class – is sufficient to determine the location of the maximum-margin hyperplane.

Consider two points, $\mathbf{x}_1 \in \mathcal{C}_1$ ($y_1 = +1$) and $\mathbf{x}_2 \in \mathcal{C}_2$ ($y_2 = -1$). The maximum margin hyperplane is therefore determined by solving

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{w}\|_2^2$$

subject to the constraints

$$y_1(\mathbf{w}^\mathsf{T}\mathbf{x}_1 + b) = 1 \text{ or } \mathbf{w}^\mathsf{T}\mathbf{x}_1 + b = 1$$
$$y_2(\mathbf{w}^\mathsf{T}\mathbf{x}_2 + b) = 1 \text{ or } -(\mathbf{w}^\mathsf{T}\mathbf{x}_2 + b) = 1$$

To solve this constrained optimization problem, we introduce Lagrange multipliers λ_1 and λ_2.

$$\arg \min_{\mathbf{w},b} \left\{ \frac{1}{2}\|\mathbf{w}\|_2^2 + \lambda_1 \left(\mathbf{w}^\mathsf{T}\mathbf{x}_1 + b - 1\right) - \lambda_2 \left(\mathbf{w}^\mathsf{T}\mathbf{x}_2 + b + 1\right) \right\}$$

To find $\mathbf{w}$ we can take the derivative with respect to $\mathbf{w}$ and set it equal to zero

$$\frac{\partial}{\partial \mathbf{w}} \left\{ \frac{1}{2}\|\mathbf{w}\|_2^2 + \lambda_1 \left(\mathbf{w}^\mathsf{T}\mathbf{x}_1 + b - 1\right) - \lambda_2 \left(\mathbf{w}^\mathsf{T}\mathbf{x}_2 + b + 1\right) \right\} = 0$$
$$0 = \mathbf{w} + \lambda_1\mathbf{x}_1 - \lambda_2\mathbf{x}_2$$

Therefore, $\mathbf{w} = -(\lambda_1\mathbf{x}_1 - \lambda_2\mathbf{x}_2)$. If we take the derivative with respect to b, and set it equal to zero we see further that

$$\frac{\partial}{\partial b} \left\{ \|\mathbf{w}\|_2^2 + \lambda_1 \left(\mathbf{w}^\mathsf{T}\mathbf{x}_1 + b - 1\right) - \lambda_2 \left(\mathbf{w}^\mathsf{T}\mathbf{x}_2 + b + 1\right) \right\} = 0$$
$$0 = \lambda_1 - \lambda_2 \tag{1}$$

Hence, $\lambda = \lambda_1 = \lambda_2$, so $\mathbf{w} = \lambda(\mathbf{x}_1 - \mathbf{x}_2)$. Note that we are going to let the negative sign be absorbed into the constant λ since the sign is rather arbitrary here. We may now solve for b, starting with our

constraints,

$$
\begin{aligned}
0 = 1 - 1 &= (\mathbf{w}^\mathsf{T}\mathbf{x}_1 + b) + (\mathbf{w}^\mathsf{T}\mathbf{x}_2 + b) \\
\rightarrow b &= -\frac{1}{2}\mathbf{w}^\mathsf{T}(\mathbf{x}_1 + \mathbf{x}_2) \\
&= -\frac{1}{2}\lambda(\mathbf{x}_1 - \mathbf{x}_2)^\mathsf{T}(\mathbf{x}_1 + \mathbf{x}_2) \\
&= -\frac{\lambda}{2}(\|\mathbf{x}_1\|_2^2 - \|\mathbf{x}_2\|_2^2) \qquad (2)
\end{aligned}
$$

We can also eliminate λ by combining our constraints differently.

$$
\begin{aligned}
2 &= (\mathbf{w}^\mathsf{T}\mathbf{x}_1 + b) - (\mathbf{w}^\mathsf{T}\mathbf{x}_2 + b) \\
&= \mathbf{w}^\mathsf{T}(\mathbf{x}_1 - \mathbf{x}_2) \\
&= \lambda(\mathbf{x}_1 - \mathbf{x}_2)^\mathsf{T}(\mathbf{x}_1 - \mathbf{x}_2) = \lambda\,\sigma(\mathbf{x}_1 - \mathbf{x}_2) \\
\lambda &= \frac{2}{\sigma(\mathbf{x}_1 - \mathbf{x}_2)}
\end{aligned}
$$

2 Linear Regression with Regularization [10pts]

In class we derived and discussed linear regression in detail. Find the result of minimize the loss of sum of the squared errors; however, add in a penalty for an L_2 penalty on the weights. More formally,

$$\arg\min_{\mathbf{w}} \left\{ \sum_i (\mathbf{w}^\mathsf{T}\mathbf{x}_i - y_i)^2 + \lambda \|\mathbf{w}\|_2^2 \right\}$$

How does this change the solution to the original linear regression solution? What is the impact of adding in this penalty?

Solution Follow the same steps we used when we discussed linear regression.

$$\begin{aligned}
\frac{\mathsf{d}}{\mathsf{d}\mathbf{w}} \left\{ \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^\mathsf{T}\mathbf{w} \right\} &= -\mathbf{X}^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w} = 0 \\
-\mathbf{X}^\mathsf{T}\mathbf{y} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} + \lambda I\mathbf{w} &= 0 \\
(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda I)\mathbf{w} &= \mathbf{X}^\mathsf{T}\mathbf{y} \\
\mathbf{w} &= (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda I)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}
\end{aligned}$$

Write your own implementation of logistic regression and implement your model on either real-world (see Github data sets: https://github.com/gditzler/UA-ECE-523-Sp2018/tree/master/data), or synthetic data. If you simply use Scikit-learn's implementation of the logistic regression classifier, then you'll receive zero points. A full 10/10 will be awarded to those that implement logistic regression using the optimization of cross-entropy using stochastic gradient descent.

3 Density Estimation [20pts]

The ECE523 Lecture notes has a function for generating a checkerboard data set. Generate checkerboard data from two classes and use any density estimate technique we discussed to classify new data using

$$\widehat{p}_{Y|X}(y|x) = \frac{\widehat{p}_{X|Y}(x|y)\widehat{p}_Y(y)}{\widehat{p}_X(x)}$$

where $\widehat{p}_{Y|X}(y|x)$ is your estimate of the posterior given you estimates of $\widehat{p}_{X|Y}(x|y)$ using a density estimator and $\widehat{p}_Y(y)$ using a maximum likelihood estimator. You should plot $\widehat{p}_{X|Y}(x|y)$ using a pseudo color plot (see https://goo.gl/2SDJPL). Note that you must model $\widehat{p}_X(x)$, $\widehat{p}_Y(y)$, and $\widehat{p}_{X|Y}(x|y)$. Note that $\widehat{p}_X(x)$ can be calculated using the Law of Total Probability.

4 Conceptual [5pts]

The Bayes decision rule describes the approach we take to choosing a class ω for a data point $\mathbf{x}$. This can be achieved modeling $P(\omega|\mathbf{x})$ or $P(\mathbf{x}|\omega)P(\omega)/P(\mathbf{x})$. Compare and contrast these two approaches to modeling and discuss the advantages and disadvantages. For the latter model, why might knowing $P(\mathbf{x})$ be useful?

Solution This question is asking use to consider the tradeoffs between a generative and discriminative classifier. A discriminative classifier attempts to model $\mathbb{P}(\omega|\mathbf{x})$, which is typically easier to estimate since we done not make assumptions or directly try to model the distribution of the data. A generative classifier, while generally being a more complex model, allows us to directly model the distribution of the data. On top of obtaining the Bayes classifier, we can directly compute $\mathbb{P}(\mathbf{x})$. This term is useful for looking for outliers in the data.