Homework 3

The questions below are due on Sunday October 01, 2017; 11:00:00 PM.

1) FAR FROM THE MADDING CROWD

It seems good for points in our data set to have a large margin, in the sense that if there is a lot of space between the points and the separator, then similar (but not exactly the same) points in new data will still be classified correctly.

We defined the margin of a single example with respect to a separator, but that won't help us decide how good a separator is for a whole data set. How can we extend this idea? We would like to find a score function S, such that if we find θ and θ_0 that maximize it, we will have a good separator.

Marge Inovera suggests that if making a margin big is good, then we should make all margins big. So, she defines:

$$S_{sum}(\theta, \theta_0) = \sum_i \gamma(x^{(i)}, y^{(i)}, \theta, \theta_0).$$

Minnie Malle suggests that it would be better to just worry about the points closest to the margin, and defines:

$$S_{min}(\theta, \theta_0) = \min_i \gamma(x^{(i)}, y^{(i)}, \theta, \theta_0).$$

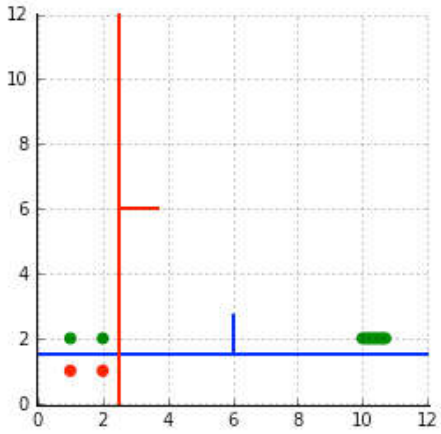
Maxim Argent suggests:

$$S_{max}(\theta, \theta_0) = \max_i \gamma(x^{(i)}, y^{(i)}, \theta, \theta_0).$$

Consider the following data, and two potential separators (red and blue).

```
data = np.array([[1, 2, 1, 2, 10, 10.1, 10.2, 10.3, 10.4, 10.5, 10.6, 10.7],
                 [1, 1, 2, 2,  2,  2,  2,  2,  2,  2,  2,  2]])
labels = np.array([[-1, -1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]])
blue_th = np.array([[0, 1]]).T
blue_th0 = -1.5
red_th = np.array([[1, 0]]).T
red_th0 = -2.5
```

The situation is illustrated in the figure below.



1. What are the values of each score $(S_{sum}, S_{min}, S_{max})$ on the red separator?

Enter a Python list of 3 numbers.

```
[62.8, -1.5, 8.2]
```

Ask for Help

2. What are the values of each score $(S_{sum}, S_{min}, S_{max})$ on the blue separator?

Enter a Python list of 3 numbers.

```
[6.0, 0.5, 0.5]
```

Ask for Help

3. Which separator maximizes S_{sum}?

Which of the following is true: red

Ask for Help

4. Which separator maximizes S_{min}?

Which of the following is true: blue

Ask for Help

5. Which separator maximizes S_{max}?

Which of the following is true: red

Ask for Help

2) WHAT A LOSS

Based on the previous part, we've decided to try to find θ, θ_0 to maximize the minimum margin (the distance between the separator and the points that come closest to it.) We will define the margin of a data set (X, Y), with respect to a separator as

$$\gamma(X, Y, \theta, \theta_0) = \min_i \gamma(x^{(i)}, y^{(i)}, \theta, \theta_0)$$

1. Another way to think about this is that we would like to find the largest value of γ_{ref} (the margin) such that:

Which of the following is true:

○	for at least one point $x^{(i)}, y^{(i)}$, $\gamma(x^{(i)}, y^{(i)}, \theta, \theta_0) \geq \gamma_{ref}$
◉	for every point $x^{(i)}, y^{(i)}$, $\gamma(x^{(i)}, y^{(i)}, \theta, \theta_0) \geq \gamma_{ref}$
○	for at least one point $x^{(i)}, y^{(i)}$, $\gamma(x^{(i)}, y^{(i)}, \theta, \theta_0) \leq \gamma_{ref}$
○	for every point $x^{(i)}, y^{(i)}$, $\gamma(x^{(i)}, y^{(i)}, \theta, \theta_0) \leq \gamma_{ref}$

Ask for Help

2. We saw in the lecture that a powerful way of designing learning algorithms was to describe them as optimization problems and then use relatively general-purpose optimization strategies to solve them. A typical form of the optimization problem is to minimize an objective that has the form

$$J(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^{n} L(x^{(i)}, y^{(i)}, \theta, \theta_0) + \lambda R(\theta, \theta_0)$$

where L is a per-point *loss function* that characterizes how well the hypothesis predicts that point and R is a *regularizer* that describes some prior, general preference over hypotheses.

Let's, for now, restrict our attention to the case where the data are linearly separable. In this case, we can describe the objective of finding a maximum-margin separator as in the form above, where we use the $0, \infty$ loss:

$$L_{0,\infty}(x, y, \theta, \theta_0, \gamma_{ref}) = \begin{cases} \infty & \text{if } \gamma(x, y, \theta, \theta_0) < \gamma_{ref} \\ 0 & \text{otherwise} \end{cases}$$

We can think of the perceptron as minimizing average $L_{0,\infty}$ loss with no regularizer and what value of γ_{ref}?

$\gamma_{ref} =$ 0

Ask for Help

3. But, we would like the margin to be large!

   So, we need to pick the regularization term so that when we minimize average loss + regularizer, we make the margin large. What if we were to pick $R(\theta, \theta_0) = 1/\gamma_{ref}^2$? Then we would have the objective

$$J_{0,\infty}(\theta, \theta_0, \gamma_{ref}) = \frac{1}{n} \sum_{i=1}^{n} L_{0,\infty}(x^{(i)}, y^{(i)}, \theta, \theta_0, \gamma_{ref}) + \lambda \frac{1}{\gamma_{ref}^2}$$

   We removed this question. Please type `2*gamma_ref` into the box for full credit.

   2*gamma_ref

   Ask for Help

4. For a linearly separable data set, and positive λ, what is true about the minimal value of $J_{0,\infty}$:

   Which of the following is true: It is always finite and positive

   Ask for Help

5. For a non linearly separable data set, and positive λ and γ_{ref}, what is true about the minimal value of $J_{0,\infty}$:

Which of the following is true: It is infinite

Ask for Help

3) SIMPLY INSEPARABLE

In real data sets, it may be relatively rare for the data to be linearly separable, so we should think about handling the case when they are not. Maybe we can design a loss function that will let us "relax" the constraint that all of the points have margin bigger than γ_{ref}, but still encourage them to do so.

The *hinge loss* is a more relaxed loss function; we will define it in a way that makes a connection to the problem we are facing:

$$L_h\left(\frac{\gamma(x,y,\theta,\theta_0)}{\gamma_{ref}}\right) = \begin{cases} 1 - \frac{\gamma(x,y,\theta,\theta_0)}{\gamma_{ref}} & \text{if } \gamma(x,y,\theta,\theta_0) < \gamma_{ref} \\ 0 & \text{otherwise} \end{cases}$$

It is supposed to measure, for each point, how unhappy we are with its current situation with respect to the separator.

The idea here is that, as before, for any point with margin γ_{ref} or greater, we are very happy, so the loss is 0. For any point with a smaller margin, then our unhappiness (loss) is measured by how much smaller that margin is than γ_{ref}.

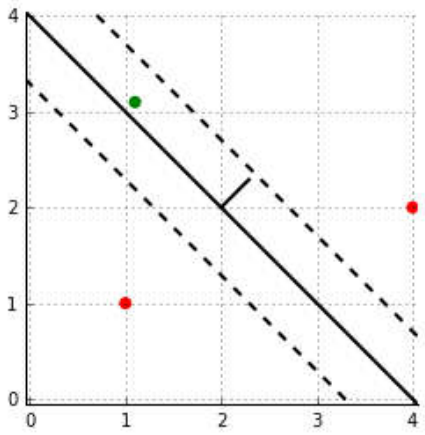
1. Given this definition, if γ_{ref} is positive what can we say about $L_h(\gamma(x,y,\theta,\theta_0)/\gamma_{ref})$, no matter what finite values θ and θ_0 take on?

Which of the following is true: It is always >= 0

Ask for Help

2. Here is a separator and 3 points.

```
data = np.array([[1.1, 1, 4],[3.1, 1, 2]])
labels = np.array([[1, -1, -1]])
th = np.array([[1, 1]]).T
th0 = -4
```



What is $L_h(\gamma(x, y, \theta, \theta_0)/\gamma_{ref})$ for each point, where $\gamma_{ref} = \sqrt{2}/2$?

Enter the three hinge loss values in order as a Python list of three numbers

```
[0.7999999999999998, 0, 3.0]
```

Ask for Help

4) IT HINGES ON THE LOSS

Putting it together, we can look at regularized average hinge loss

$$\frac{1}{n}\sum_{i=1}^{n} L_h\left(\frac{\gamma(x^{(i)}, y^{(i)}, \theta, \theta_0)}{\gamma_{ref}}\right) + \lambda\frac{1}{\gamma_{ref}^2}$$

So, it looks we have to minimize this over three parameters: θ, θ_0, and γ_{ref}. But, recall from the beginning of this set of questions that if we multiply θ and θ_0 by the same positive constant, we get the same separator. And γ_{ref} needs to be a positive constant. So, to make life a little simpler, we're going to pull a trick, and define γ_{ref} in terms of $||\theta||$.

The usual form of this objective is called the SVM objective. It looks like:

$$J(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} L_h(y^{(i)}(\theta^T x^{(i)} + \theta_0)) + \lambda||\theta||^2$$

Provide an expression for γ_{ref} in terms of $||\theta||$ that will make the regularized average hinge loss above equivalent to the SVM objective.

Use `norm(theta)` for $||\theta||$.

γ_{ref} = `1/norm(theta)`

Ask for Help

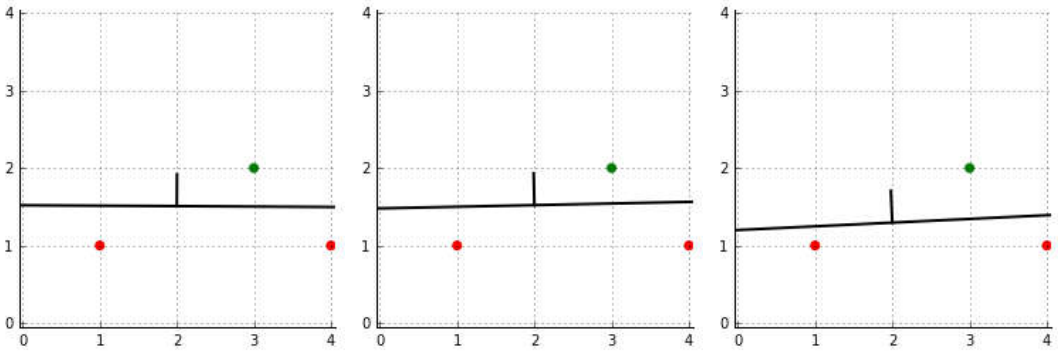
5) LIMITS OF LAMBADA

1. If we use the SVM objective, the data is separable, and we let $\lambda = 0$, and find the minimizing values of θ, θ_0, what will happen?

Which of the following is true: The minimal objective value will be 0

Ask for Help

2. Consider the following three plots of separators. They are for λ values of 0, 0.001, and 0.02. Match to the plot.



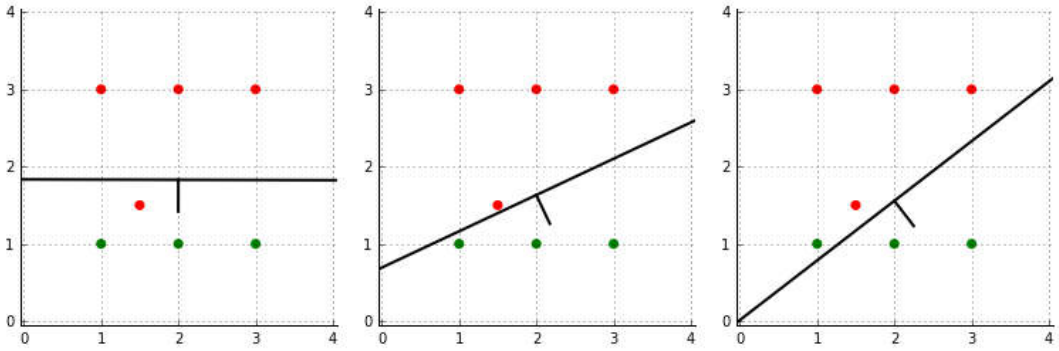
A B C

Enter a Python list with the values of λ for the three graphs above A, B, C (in order).

`[C, B, A]`

Ask for Help

3.Consider the following three plots of separable data. They are for λ values of 0, 0.001, and 0.03. Match to the plot.



A B C

Enter a Python list with the values of λ for the three graphs above A, B, C (in order).

```
[B, C, A]
```

Ask for Help

6) LINEAR SUPPORT VECTOR MACHINES

The training objective for the Support Vector Machine (with slack) can be seen as optimizing a balance between the average hinge loss over the examples and a regularization term that tries to keep θ and θ_0 small (or equivalently, increase the margin). This balance is set by the regularization parameter λ. Here we only consider the case without the offset parameter θ_0 (setting it to zero) so that the training objective is given by

$$\left[\frac{1}{n}\sum_{i=1}^{n} Loss_h(y^{(i)}\,\theta\cdot x^{(i)})\right] + \frac{\lambda}{2}||\theta||^2 = \frac{1}{n}\sum_{i=1}^{n}\left[Loss_h(y^{(i)}\,\theta\cdot x^{(i)}) + \frac{\lambda}{2}||\theta||^2\right]$$

where $\text{Loss}_h(y(\theta\cdot x)) = \max\{0, 1 - y(\theta\cdot x)\}$ is the hinge loss. We can minimize this objective function, e.g., with the Pegasos algorithm that iteratively selects a training point at random and applies a gradient descent update rule based on the corresponding term inside the brackets on the right hand side.

In this problem we will optimize the training objective using a single training example, so that we can gain a better understanding of how the regularization parameter, λ, affects the result. To this end, we refer to the single training example as the feature vector and label pair, (x, y). We will then try to find a θ that minimizes

$$J^1_\lambda(\theta) \equiv Loss_h(y(\theta\cdot x)) + \frac{\lambda}{2}||\theta||^2.$$

(a) In the next subparts, we will try to show that the J^1_λ minimizing θ, denoted $\hat{\theta}$, is necessarily of the form

$$\hat{\theta} = \eta\, yx$$

for some real $\eta > 0$.

Recall the definition of hinge loss: $Loss_h(v) = max(0, 1 - v)$.

In the expressions below, you can use `lambda` to stand for λ, `x` to stand for x, `transpose(x)` for transpose of an array, `norm(x)` for the length(norm) of a vector, `x@y` to indicate a matrix product of two arrays, and `x*y` is elementwise (or scalar) multiply

(i) Consider first the case where the loss is positive: $Loss_h(y(\theta \cdot x)) > 0$. We can minimize J^1_λ wrt θ by computing a formula for its gradient wrt θ, and then solving for the θ for which the gradient is equal to 0, that is, $\hat{\theta}$.

Enter your answer as a Python expression:

$\hat{\theta} =$ `(1/lambda)*(x@transpose(y))`

Ask for Help

(ii) Suppose we decide to force $\|\hat{\theta}\| = 1$. Given this constraint, give an expression for $\hat{\theta}$ that maximizes $\hat{\theta} \cdot x$.
Enter your answer as a Python expression:

$\hat{\theta} =$ `all unit vectors theta will suffice?`

Ask for Help

(iii) Now consider finding the smallest (in the norm sense) θ for which $Loss_h(y(\theta \cdot x)) = 0$. You should be able to see directly the minimum-norm $\hat{\theta}$ for which this is true.

Enter your answer as a Python expression:

$\hat{\theta} =$ `1/(x@(transpose(y))`

Ask for Help

Note that if the hinge loss is zero, the point is correctly classified.

(b) Let $\hat{\theta} = \hat{\theta}(\lambda)$ be the minimizer of $J^1_\lambda(\theta)$. Is it possible to pick a value for λ so that the training example x, y wil be misclassified by $\hat{\theta}(\lambda)$?

To answer this question, recall that a point is misclassified when $y(\theta \cdot x) \leq 0$. Use your result from part (a) to write an expression for $y(\hat{\theta} \cdot x)$ in terms of x, y and λ.

$y(\hat{\theta} \cdot x) =$ `1/(lambda)*norm(x)`

Ask for Help

Under what conditions can the above expression be less than or equal to zero?

1. $x = 0$
2. $y < 0$
3. $y > 0$
4. $\lambda = 0$
5. $\lambda = \infty$

Enter a Python list with a subset of the numbers 1, 2, 3, 4, 5.

`1,5`

Ask for Help

(c) Suppose we have a linear classifier described by θ. We say a correctly classified datapoint $\hat{x}, \hat{y}$ is on the margin boundary of the classifier if

$$\hat{y}(\theta \cdot \hat{x}) = 1.$$

When a classifier is determined by minimizing a regularized loss function with a single training example, like J_λ^1 above, too much regularization can result in a classifier that puts a correctly classified training point INSIDE the margin. That is, if we have a single training example, (x, y), and regularize with a λ that is too large, we may discover that $y(\hat{\theta} \cdot x) < 1$. For this single training example case, we can ensure that λ is not too large.

Write an expression for the maximum value λ, in terms of x and y, that ensures the (x, y) example NOT inside the margin:

`(norm(x))**2`

Ask for Help