Hull-Robert-HW2-final.md

- Robert 'Quinn' Hull

- 02/17/2021
- HW 2

Outline

1. **Linear Classifier with a Margin (10 pt)**
2. **Linear Regression with Regularization (10 pts)**
3. **Density Estimation (20 pts)**
4. **Conceptual (5 pts)**

1. Linear Classifier with a Margin (10 pt)

1 Linear Classifier with a Margin [10pts]

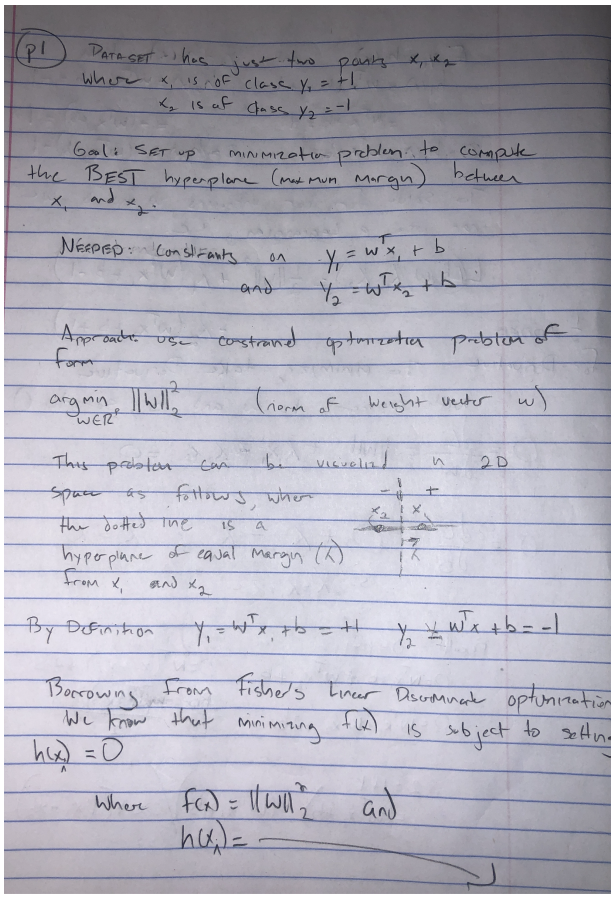
Show that, regardless of the dimensionality of the feature vectors, a data set that has just two data points, one from each class, is sufficient to determine the location of the maximum-margin hyperplane. Hint #1: Consider a data set of two data points, $\mathbf{x}_1 \in \mathcal{C}_1$ ($y_1 = +1$) and $\mathbf{x}_2 \in \mathcal{C}_2$ ($y_2 = -1$) 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.

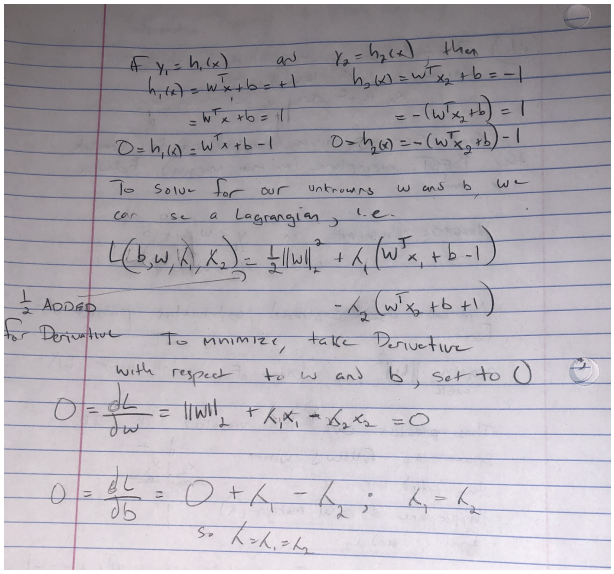
$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \|\mathbf{w}\|_2^2$$

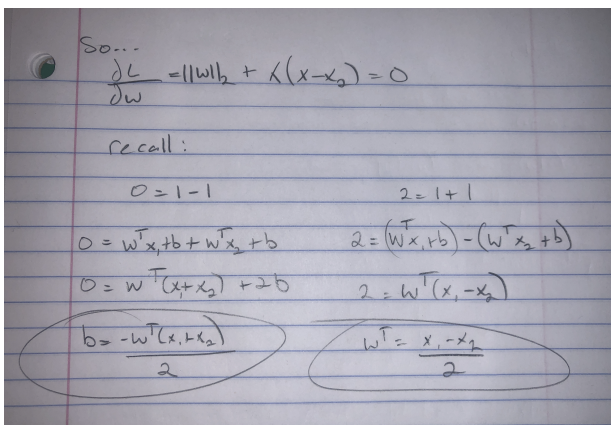
Subject to: (some constraint)

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)?

Answer


P1
DATASET: has just two points x_1, x_2
where x_1 is of class $y_1 = +1$
x_2 is of class $y_2 = -1$
Goal: SET up minimization problem to compute the BEST hyperplane (maximum margin) between x_1 and x_2.
NEEDED: constraints on $y_1 = w^T x_1 + b$
and $y_2 = w^T x_2 + b$
Approach: use constrained optimization problem of form
$\arg\min_{w \in \mathbb{R}^n} \|w\|_2^2$ (norm of weight vector w)
This problem can be visualized in 2D space as follows, where the dotted line is a hyperplane of equal margin (λ) from x_1 and x_2
By Definition $y_1 = w^T x_1 + b = +1$ $y_2 = w^T x + b = -1$
Borrowing from Fisher's Linear Discriminant optimization we know that minimizing $f(x)$ is subject to setting $h(x_\lambda) = 0$
where $f(x) = \|w\|_2^2$ and
$h(x_\lambda) =$


If $y_1 = h_1(x)$ and $y_2 = h_2(x)$ then
$h_1(x) = w^T x_1 + b = +1$
$= w^T x + b = 1$
$0 = h_1(x) = w^T x + b - 1$
$h_2(x) = w^T x_2 + b = -1$
$= -(w^T x_2 + b) = 1$
$0 = h_2(x) = -(w^T x_2 + b) - 1$
To solve for our unknowns w and b, we can set a Lagrangian, i.e.
$L(b, w, \lambda_1, \lambda_2) = \frac{1}{2}\|w\|_2^2 + \lambda_1(w^T x_1 + b - 1)$
$- \lambda_2(w^T x_2 + b + 1)$
$\frac{1}{2}$ ADDED for Derivative
To minimize, take Derivative with respect to w and b, set to 0
$0 = \frac{\partial L}{\partial w} = \|w\|_2 + \lambda_1 x_1 - \lambda_2 x_2 = 0$
$0 = \frac{\partial L}{\partial b} = 0 + \lambda_1 - \lambda_2 \therefore \lambda_1 = \lambda_2$
So $\lambda = \lambda_1 = \lambda_2$


So...
$\frac{\partial L}{\partial w} = \|w\|_2 + \lambda(x - x_2) = 0$
recall:
$0 = 1 - 1$
$2 = 1 + 1$
$0 = w^T x_1 + b + w^T x_2 + b$
$2 = (w^T x_1 + b) - (w^T x_2 + b)$
$0 = w^T(x_1 + x_2) + 2b$
$2 = w^T(x_1 - x_2)$
$b = \frac{-w^T(x_1 + x_2)}{2}$
$w^T = \frac{x_1 - x_2}{2}$

2. Linear Regression with Regularization (10 pts)

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,

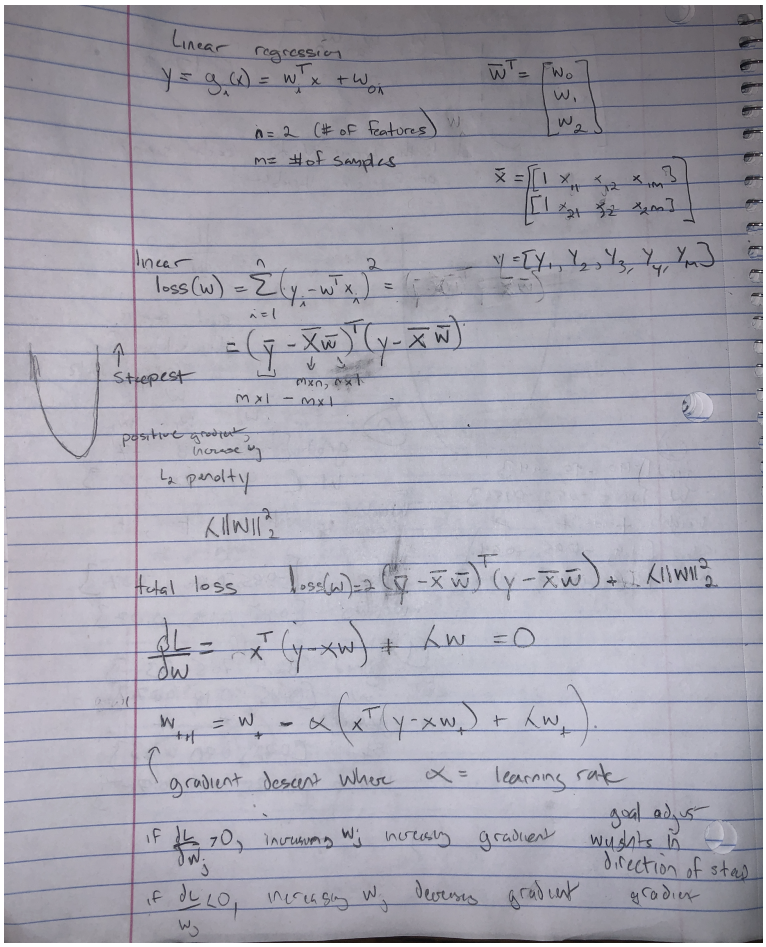
$$\arg\min_{\mathbf{w}} \left\{ \sum_i (\mathbf{w}^\mathsf{T}\mathbf{x}_i - y_i)^2 + \lambda \|\mathbf{w}\|_2^2 \right\}$$

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

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.

Answer

NOTE COMPLETE IMPLEMENTATION IN CODE BELOW



3. Density Estimation (20 pts)

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.

Answer

NOTE COMPLETE IMPLEMENTATION IN CODE BELOW

4. Conceptual (5 pts)

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?

Answer:

- The scribe on discriminant functions provides a good overview for this question that response mirrors, thank you!
- The question asks how we can use probability to classify discrete data using a Bayes Decision Rule framework that minimizes the probability of a classification error. Namely:

> $$\omega^* = \arg\max_{\omega\in\Omega} p(\mathbf{x}|\omega)p(\omega) = \arg\max_{\omega\in\Omega} p(\omega|\mathbf{x})$$

> where `w*` is the predicted class, `p(x|w)` is the likelihood, `p(w)` is the prior, and `p(w|x)` is the posterior probability.

- The left-hand-side (LHS) approach to modeling `w*` requires estimating the probability of multiple quantities: likelihood, prior, and (implicitly) evidence (`p(x)`). These are called '**generative**' models. Quoting directing from the scribe, *'this is the equivalent to attempting to directly estimate the joint distribution `p(x,w)` and normalizing to obtain the posterior probabilities'*
  - An advantage of **generative** models is that they can give you more information. In particular, because you know information about likelihood, priors, and evidence, you can use generative models to generate synthetic data that fit a particular distribution. This can be really powerful for creating reproducible, interperable results.
  - A disadvantage is that determining all of these characteristics of the data (likelihoods and priors) is empirically difficult in many situations, particularly with high-dimensional data where these computations are expensive. For example, calculating the likelihood (`px|w`) in a classification problem involving fotos, which can have 1000x1000 pixels, is in practice very difficult, especially for very large datasets, because this calculation has to be done for every single pixel.
- The right-hand-side (RHS) approach to modeling `w*` requires estimating just the posterior `p(w|x)` directly. This is called a **discriminative** model.
  - An advantage of this is that the calculations are usually easier, because you aren't trying to calculate three different quantities in your decision.

    - A disadvantage is that we know nothing of the joint distribution `p(w, x)`, which makes our estimation a little less informative, and perhaps more uncertain. Since we are unable to to estimate the likelihood, we can not generate data, or detect outliers.
    - I'm guessing, but I think it might be nice to know `p(x)` for discriminative models because then we can directly compare the quantities calculated from the discriminative and generative approaches. i.e. `p(w|x)*p(x) = p(x:w)*p(w)`, according to Bayes Decision Rule.

- My final comment on this is that the approach described in Problem 2 (linear regression) is neither generative nor discriminative, and instead a **discriminant function**. The text establishes that this approach doesn't use a likelihood based approach, and thus makes prediction without having any probabilistic confidence.