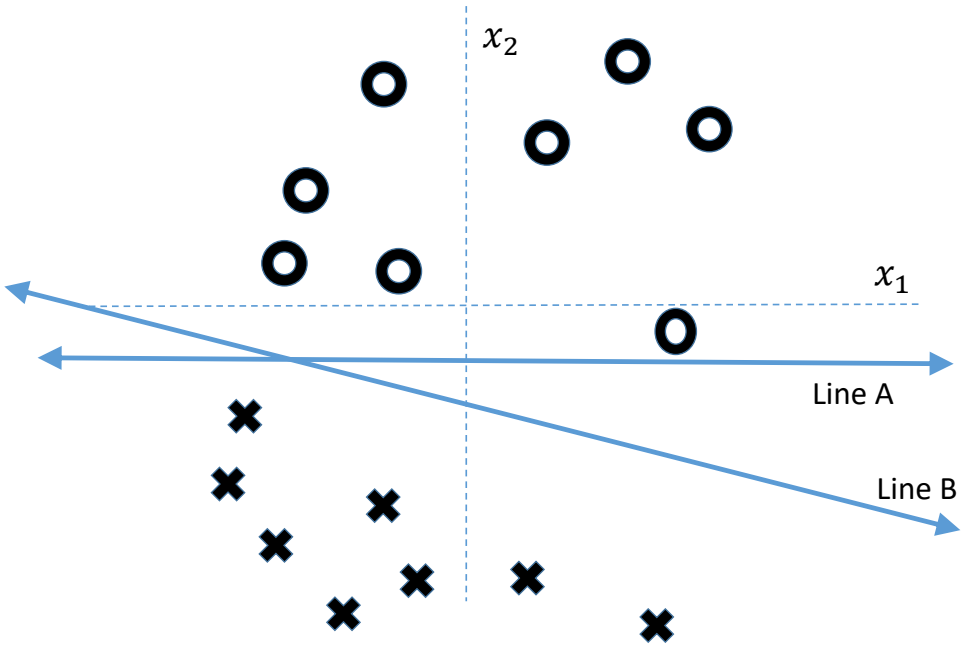
Section 05: Midterm Review

1. Logistic Regression

Let's think about what happens with linearly separable data in logistic regression.

Consider the following training set:



Each data point has two features – one shown on the horizontal axis, and the second on the vertical axis. Each point is in one of two classes – 'X' is class 1, 'O' is class -1. Our model is logistic regression:

$$\Pr(y = 1|x, w) = \frac{1}{1 + \exp(-w_0 - w_1 x[1] - w_2 x[2])}$$

We've shown two lines that might have been the result of training on our training set.

(a) What are some possible values of w to produce line A and line B. Line A is horizontal with an intercept of -1. Line B has a slope of $-1/2$ and an intercept of -2 on the vertical axis.

(b) Show that for the w you found in part a, that $2w$ still corresponds to the same separating hyperplane (i.e. to the same lines A and B)

(c) In lecture, we said a common loss function for this problem is $J(w) = \sum_{i=1}^{n} \log\left(1 + \exp(-y_i x_i^T w)\right)$ If we try to minimize this error, the w we found in part a. won't actually be $\hat{w} = \arg\min \sum_{i=1}^{n} \log\left(1 + \exp(-y_i x_i^T w)\right)$ – why? (think about part b)

(d) Why is this an issue, and how do we fix it?

(e) Suppose we regularized our error, by adding the term $\lambda(|w_1| + |w_2|)$ to our objective (i.e. L1 regularization). For large values of λ which of the lines do you expect to have smaller objective value (i.e. smaller error plus regularization penalty)? (We're looking for intuition, not a calculation. Hint: why do we usually use L1 regularization?)

(f) We didn't regularize by w_0. Give an example of a data set where regularizing with $\lambda(|w_0| + |w_1| + |w_2|)$ would lead to much worse behavior than regularizing by $\lambda(|w_1| + |w_2|)$.

(g) Suppose you've let λ grow so large that w_1 and w_2 have been set to 0. What kind(s) of functions are we now capable of modeling? What value of w_0 will minimize the error?

2. Solving Problems

You're preparing for a meeting with a client that you just finished making a regression model for, and the client does NOT seem happy. You have the ability to do any of the following techniques to try to improve your model.

a. Find a larger training set.

b. Try using a more complex hypothesis class

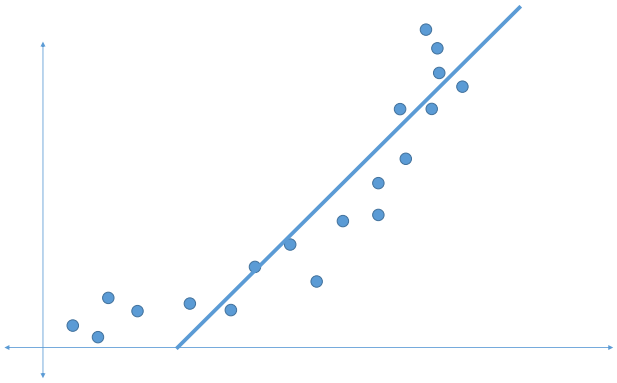
c. Try L1 regularization

d. Change the set of features

For each of the following complaints your client could give you, which of the above techniques is most likely to make them happy?

(a) "I can't tell what this thing is doing. Is the number of bathrooms important for the selling price or not?"

(b) "Look at this plot, do these predictions look accurate to you?



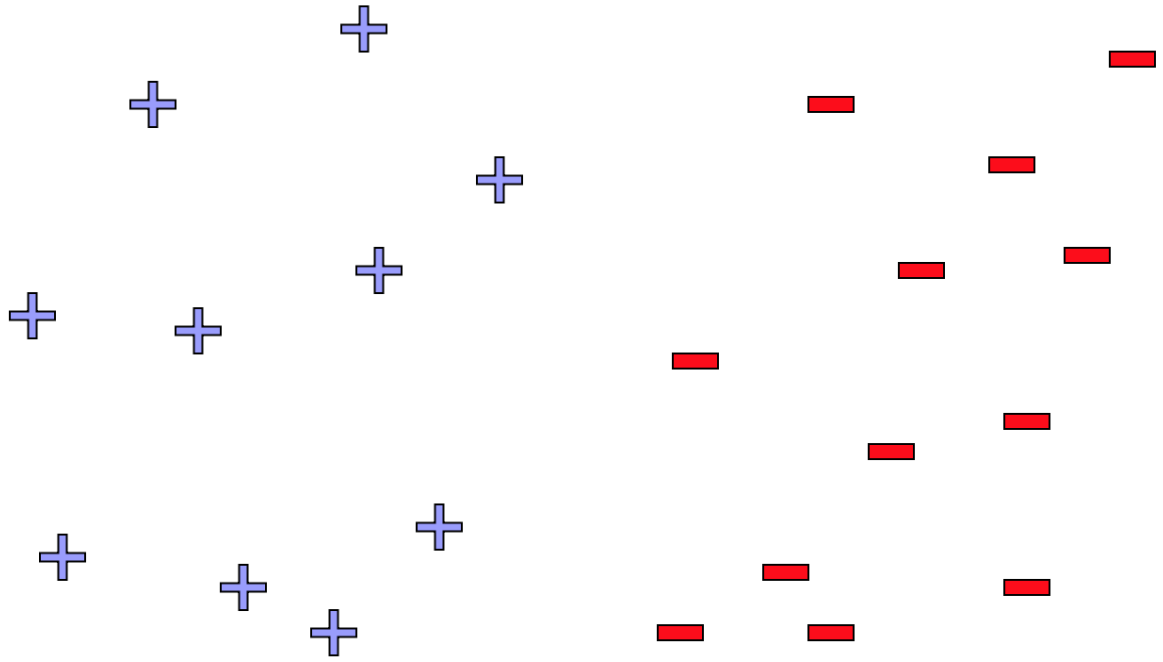
(c) "Can you really recognize handwritten digits with just linear combinations of pixel values? It doesn't seem like we're getting low error here."

(d) "We trained your model on two different subsets of the test set and got very very different predictions from each half."

3. More Practice

3.1. More Logistic Practice

Let's think about things we need to be careful about in machine learning. We will use an important concepts in machine learning to solve this problem. Consider the following picture obtained from lecture,



where each data points are i.i.d. of the form $(x_i, y_i)_{i=1}^n$ and $x_i \in R^d$, $y_i \in \{-1, 1\}$. Intuitively, you can think of this data having $d = 2$, and imagine 2 axis draw across this image, which would simplify your thought process to solve this problem.

Notice that we will use some characteristics about these data points to inform us of our decision choices when training our algorithm.

(a) Suppose we have some classification problems to solve for the above data points. Which loss function should we use to train on?

(b) Provide justification for this loss function. What is the interpretation of each part of the function? Why is regularization particularly important in this instance?

(c) Why would it be bad not to regularize?

3.2. True/False, Multiple Choice

(a) Suppose you're using L2 regularization on a least squares objective. Some value λ^* will give you the best test error among all possible λ. You train your model using a $\lambda' << \lambda^*$. Which of the following do you expect to be true about your training error?

   a. The training error for λ' will be much smaller than for λ^*

   b. The training error for λ' will be much bigger than for λ^*

   c. The training error for each will be about the same – we regularize for the effect on test error.

(b) The maximum likelihood estimator is always unbiased.

(c) Gradient descent will always converge to a global minimum on a convex function

(d) Why might it be difficult to choose a good learning rate for gradient descent with L1 regularization?

a. The gradient will have many terms, so each update step is slow.

b. Near the minimum, when the regularization term dominates, the gradient might stay at a constant magnitude.

c. The L1 regularization makes the function non-convex, and might introduce local minima.

(e) Suppose f is a convex function, g is a concave function, and h is a linear function (i.e. $h(x+y) = h(x)+h(y)$). For each of the following, say whether the function is concave, convex, or that you can't give a guarantee.

(i) $f - g$

(ii) $f + h$

(iii) $g + h$

(iv) fg

(f) If you are trying to find the minimum of a convex function and choose a learning rate small enough for convergence, which of the following is true?

a. The function value is decreasing at each iteration for both gradient descent and stochastic gradient descent.

b. The function value is decreasing at each iteration for gradient descent but possibly not for stochastic gradient descent.

c. The function value is decreasing at each iteration for stochastic gradient descent and but may not for gradient descent.

d. The function value could go up and down for both gradient descent and stochastic gradient descent, but you'll eventually arrive at the global minimum.

3.3. Free Response Questions

(a) Describe some differences between linear and logistic regression.

(b) Why might we prefer convex loss functions over non-convex ones?

(c) Why would we want to use SGD in practice instead of GD?

(d) Sort the following models by variance of the model class:

(i) Quadratic functions

(ii) Constant functions (i.e. functions that output the same prediction for all data points)

(iii) Linear functions

4. Machine Learning Terms

Below is a list of terms we've defined in class. If you aren't comfortable with the meanings of each of these, you should probably look them up in the textbook or lecture slides.

- bias, variance, irreducible error
- unbiased (estimator)
- feature
- feature map
- linear least squares
- ridge
- lasso
- L1
- L2
- Linfinity
- regularization
- regularization path
- overfit
- underfit
- model/hypothesis class
- regression
- classification
- training set
- validation set
- test set
- cross-validation
- k-fold cross validation
- logistic regression
- sigmoid