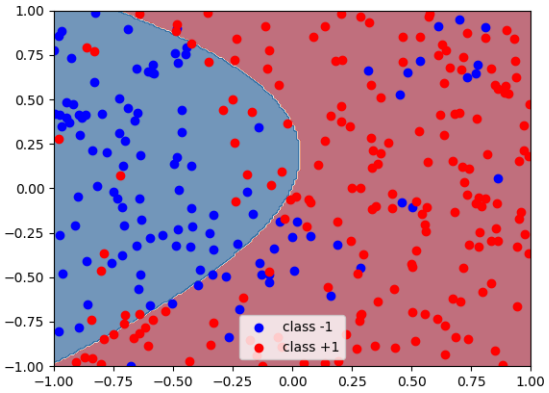
CPSC 340 Assignment 5 (due Friday November 16 at 11:55pm)

1 Kernel Logistic Regresion

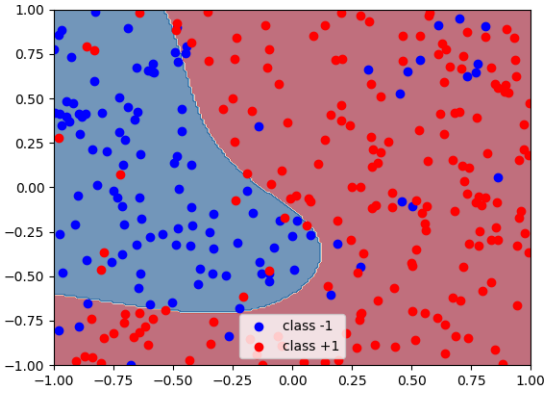
If you run `python main.py -q 1` it will load a synthetic 2D data set, split it into train/validation sets, and then perform regular logistic regression and kernel logistic regression (both without an intercept term, for simplicity). You'll observe that the error values and plots generated look the same since the kernel being used is the linear kernel (i.e., the kernel corresponding to no change of basis).

1.1 Implementing kernels

Polynomial kernel: Training error 0.183 Validation error 0.170

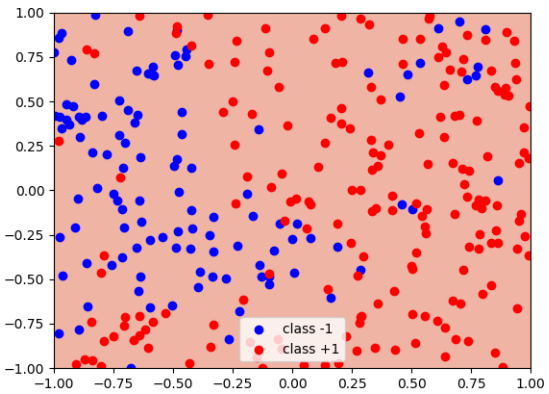


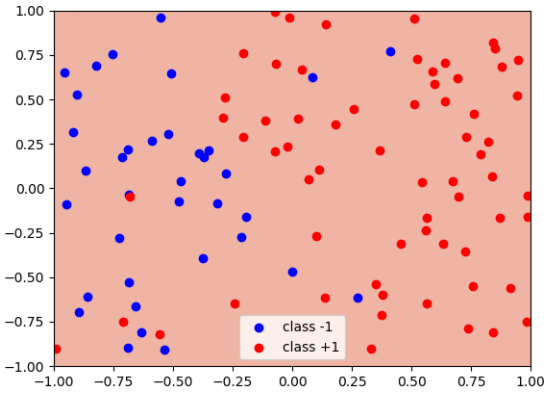
RBF Kernel: Training error 0.127 Validation error 0.090



1.2 Hyperparameter search

For both training and validation error, I got $\lambda = 1$ and $\sigma = 100$.





1.3 Reflection

Rubric: {reasoning:1}

Briefly discuss the best hyperparameters you found in the previous part, and their associated plots. Was the training error minimized by the values you expected, given the ways that σ and λ affect the fundamental tradeoff?

2 MAP Estimation

1. $f(w) = ||Xw - y||_1 + \frac{1}{2\sigma^2}||w||^2$
2. $f(w) = \frac{1}{2}(Xw - y)^T\Sigma^{-1}(Xw - y) + \lambda||w||_1$
3. The regularizer is $\frac{\lambda}{2}||w||^2$.
4.

3 Principal Component Analysis

1. We want to center the two variables around 0. x_2 is centered at 1. The centered x_2 is that all its points -1. Then, the centered variables lie along $-2x_1 = x_2$. Then the vector is (-2,1) and the magnitude is $\sqrt{(-2)^2 + 1^2} = \sqrt{5}$. Then the normalized vector is $(\frac{-2}{\sqrt{5}}, \frac{1}{\sqrt{5}})$.

2. $$z = \frac{(-3-0)}{\sqrt{5}} + \frac{(2.5-1)}{\sqrt{5}} = \frac{-3}{2\sqrt{5}}$$
$$\hat{x} = \frac{-3}{2\sqrt{5}}(\frac{-2}{\sqrt{5}}, \frac{1}{\sqrt{5}}) + (0,1) = (\frac{3}{5}, \frac{-3}{10})$$
$$\text{error} = \sqrt{(\frac{3}{5}+3)^2 + (\frac{-3}{10} - 2.5)^2} = 4.5607$$

3. $$z = \frac{(-3-0)}{\sqrt{5}} + \frac{(2-1)}{\sqrt{5}} = \frac{-2}{\sqrt{5}}$$
$$\hat{x} = \frac{-2}{\sqrt{5}}(\frac{-2}{\sqrt{5}}, \frac{1}{\sqrt{5}}) + (0,1) = (\frac{4}{5}, \frac{-2}{5})$$
$$\text{error} = \sqrt{(\frac{4}{5}+3)^2 + (\frac{-2}{5} - 2)^2} = 4.4944$$

4 PCA Generalizations

4.1 Robust PCA

Code is in pca.py and figures are in figs folder.

4.2 Reflection

Rubric: {reasoning:3}

1. Briefly explain why using the L1 loss might be more suitable for this task than L2.
2. How does the number of video frames and the size of each frame relate to n, d, and/or k?
3. What would the effect be of changing the threshold (see code) in terms of false positives (cars we identify that aren't really there) and false negatives (real cars that we fail to identify)?

5 Very-Short Answer Questions

1. The "other" normal equations are faster when $d > n$ where the cost is $O(n^2k + n^3)$ instead of $O(nk^2 + k^3)$.
2. An advantage for kernel k-means over normal k-means is that kernel k-means allows for non-convex clusters
3. MAP not only maximizes the likelihood but also the prior, $p(w)$ term which is the regularizer.
4. In generative model, we get $p(y_i|x_i)$ from using the rules of probability. For discriminitive model, $p(y_i|x_i)$ is modeled directly from fixed features.

5. No. Higher k should generally have lower loss since it contains more dimensions so the loss is minimized. In some case, higher k may result in the same loss (when all data points lie on w) but it should never be higher loss.
6. Label Switching means that the ordering of vectos minimizing PCA may be different which can lead to non-unique solutions
7. k is the number of dimensions/features we want. It wouldn't make sense for $k > d$ because then we would want more features than we already have.
8. eigenfaces will be stored in X
9. Stochastic Descent has a cost of $O(k)$ per iteration which can be useful for large datasets but it does not enforce uniqueness.
10. $\alpha^t = 1/\sqrt{t}$.
11. Global features uses features to predict what is important to all examples while local features allow us to predict what is important to a specific example.