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

Instructions

Rubric: {mechanics:5}

The above points are allocated for following the general homework instructions.

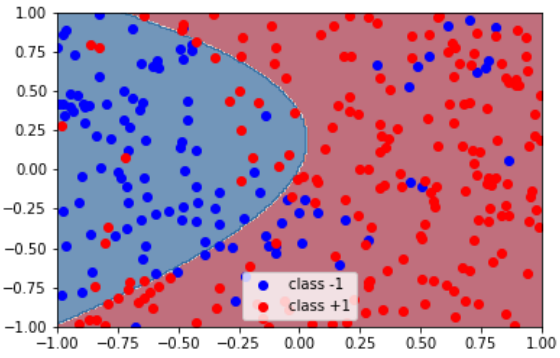
1 Kernel Logistic Regresion

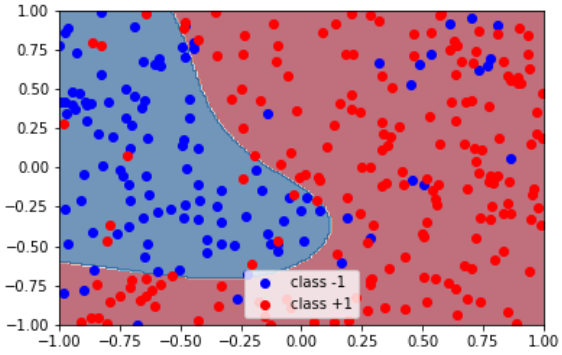
If you run `python main.py -q 1` it will load a synthetic 2D data set, split it into train/validation sets, and then perform regular logistic regression and kernel logistic regression (both without an intercept term, for simplicity). You'll observe that the error values and plots generated look the same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

1.1 Implementing kernels

Rubric: {code:5}

Implement the polynomial kernel and the RBF kernel for logistic regression. Report your training/validation errors and submit the plots generated for each case. You should use the hyperparameters $p = 2$ and $\sigma = 0.5$ respectively, and $\lambda = 0.01$ for the regularization strength. Answer: Refer to logReg.py for details. For polynomial kernel, the training error is 0.183, the validation error is 0.170 For RBC kernel, the training error is 0.127, the validation error is 0.09. The figures are shown as below.





1.2 Hyperparameter search

Rubric: {code:3}

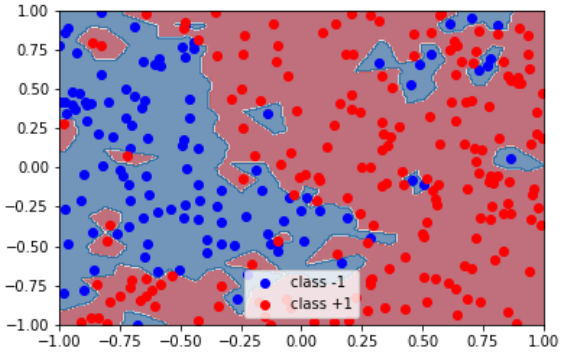
For the RBF kernel logistic regression, consider the hyperparameters values $\sigma = 10^m$ for $m = -2, -1, \ldots, 2$ and $\lambda = 10^m$ for $m = -4, -3, \ldots, 0$. In `main.py`, sweep over the possible combinations of these hyperparameter values. Report the hyperparameter values that yield the best training error and the hyperparameter values that yield the best validation error. Include the plot for each.

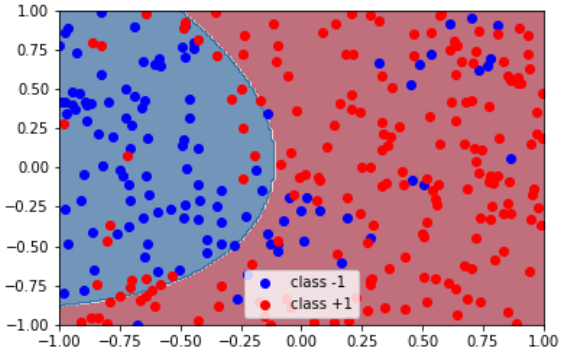
Note: on the job you might choose to use a tool like scikit-learn's `GridSearchCV` to implement the grid search, but here we are asking you to implement it yourself by looping over the hyperparameter values.

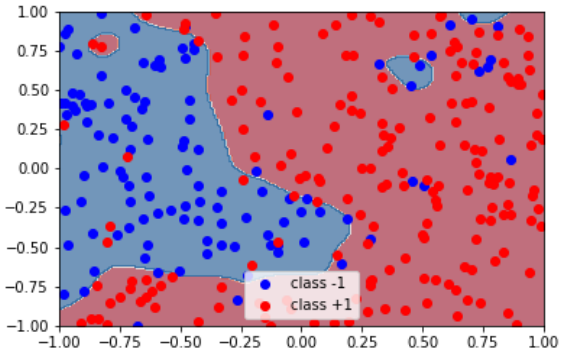
Answer:

1. The best training error is 0, the hyperparameter for σ is 0.01 and the hyperparameter for λ could be any value given between 10^{-4} and 1.
2. The best validation error is 0.12, the hyperparameter for σ is 1 and the hyperparameter for λ is 1. Or $\sigma = 0.1$ and $\lambda = 1$.

The figures are shown as follows:







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?

Answer: large λ increases training error but decreases the validation error. Large σ is to make a simple model, which may increase the training error but decrease the validation error. So the results are as expected.

2 MAP Estimation

Rubric: {reasoning:8}

In class, we considered MAP estimation in a regression model where we assumed that:

- The likelihood $p(y_i \mid x_i, w)$ is a normal distribution with a mean of $w^T x_i$ and a variance of 1.
- The prior for each variable j, $p(w_j)$, is a normal distribution with a mean of zero and a variance of λ^{-1}.

Under these assumptions, we showed that this leads to the standard L2-regularized least squares objective function,

$$f(w) = \frac{1}{2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2,$$

which is the negative log likelihood (NLL) under these assumptions (ignoring an irrelevant constant). For each of the alternate assumptions below, show how the loss function would change (simplifying as much as possible):

1. We use a Laplace likelihood with a mean of $w^T x_i$ and a scale of 1, and we use a zero-mean Gaussian prior with a variance of σ^2.

$$p(y_i \mid x_i, w) = \frac{1}{2}\exp(-|w^T x_i - y_i|), \quad p(w_j) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{w_j^2}{2\sigma^2}\right).$$

2. We use a Gaussian likelihood where each datapoint has its own variance σ_i^2, and where we use a zero-mean Laplace prior with a vairance of λ^{-1}.

$$p(y_i \mid x_i, w) = \frac{1}{\sqrt{2\sigma_i^2\pi}}\exp\left(-\frac{(w^T x_i - y_i)^2}{2\sigma_i^2}\right), \quad p(w_j) = \frac{\lambda}{2}\exp(-\lambda|w_j|).$$

   You can use Σ as a diagonal matrix that has the values σ_i^2 along the diagonal.

3. We use a (very robust) student t likelihood with a mean of $w^T x_i$ and ν degrees of freedom, and a zero-mean Gaussian prior with a variance of λ^{-1},

$$p(y_i|x_i, w) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{(w^T x_i - y_i)^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad p(w_j) = \frac{\sqrt{\lambda}}{\sqrt{2\pi}}\exp\left(-\lambda\frac{w_j^2}{2}\right).$$

   where Γ is the "gamma" function (which is always non-negative).

4. We use a Poisson-distributed likelihood (for the case where y_i represents counts), and we use a uniform prior for some constant κ,

$$p(y_i|w^T x_i) = \frac{\exp(y_i w^T x_i)\exp(-\exp(w^T x_i))}{y_i!}, \quad p(w_j) \propto \kappa.$$

   (This prior is "improper" since $w \in \mathbb{R}^d$ but it doesn't integrate to 1 over this domain, but nevertheless the posterior will be a proper distribution.)

Answer: In the following answers, we neglect the changes of the constant part. We use log-likelihood estimation to obtain the results

1. We need to change loss function to $\|Xw - y\|_1 + \frac{1}{2\sigma^2}\|w\|^2$.
2. We need to change it to $\sum_i \frac{1}{2\sigma_i^2}(w^T x_i - y_i)^2 + \lambda\|w\|_1 = \frac{1}{2}(Xw - y)^T\Sigma^{-1}(Xw - y) + \lambda\|w\|_1$
3. For the regularization part, we need to change it to $\frac{\lambda}{2}\|w\|^2$. For the data-fitting term, we we still use negative log-likelihood, we have

$$-\log(p(y_i|x_i, w)) = constant - (-\frac{\nu+1}{2}\log(1 + \frac{(w^T x_i - y_i)^2}{\nu})) \tag{1}$$

   Therefore, for the data-fitting term, we need to change it to $\sum_i \frac{\nu+1}{2}\log(1 + \frac{(w^T x_i - y_i)^2}{\nu})$ Finally, we have $\sum_i \frac{\nu+1}{2}\log(1 + \frac{(w^T x_i - y_i)^2}{\nu}) + \frac{\lambda}{2}\|w\|^2$
4. We need to change the loss function to $-y^T Xw + \sum_i(\exp w^T x_i)$

3 Principal Component Analysis

Rubric: {reasoning:3}

Consider the following dataset, containing 5 examples with 2 features each:

x_1	x_2
-4	3
0	1
-2	2
4	-1
2	0

Recall that with PCA we usually assume that the PCs are normalized ($\|w\| = 1$), we need to center the data before we apply PCA, and that the direction of the first PC is the one that minimizes the orthogonal distance to all data points.

1. What is the first principal component?
2. What is the reconstruction loss (L2 norm squared) of the point (-3, 2.5)? (Show your work.)
3. What is the reconstruction loss (L2 norm squared) of the point (-3, 2)? (Show your work.)

Hint: it may help (a lot) to plot the data before you start this question.

Answer: We first center the data. Since the mean of the first variable is zero, we do not need to center it. For the second variable, the mean is 1. So, we have the centered data like

x_1	x_2
-4	2
0	0
-2	1
4	-2
2	-1

1. By plotting the data, we find that the line that minimizes the orthogonal distance of the data is passing $(4, -2)$, so the direction can be $(2, -1)$. After the normalization, we have $w_1 = (2/\sqrt{5}, -1/\sqrt{5})$
2. We first center the data and get $(-3, 1.5)$. We then get the low-dimensional representation as
$$z = -3 \times \frac{2}{\sqrt{5}} - 1.5 \times \frac{1}{\sqrt{5}} = -7.5/\sqrt{5} \tag{2}$$
The we get
$$\hat{x} = \frac{-7.5}{\sqrt{5}}(2/\sqrt{5}, -1/\sqrt{5}) + (0, 1) = (-3, 2.5) \tag{3}$$
So the reconstruction error is zero
3. We first center the data and get $(-3, 1)$. We then get the low-dimensional representation as
$$z = -3 \times \frac{2}{\sqrt{5}} - 1 \times \frac{1}{\sqrt{5}} = -7/\sqrt{5} \tag{4}$$
The we get
$$\hat{x} = \frac{-7}{\sqrt{5}}(2/\sqrt{5}, -1/\sqrt{5}) + (0, 1) = (-2.8, 1.4) \tag{5}$$
So the reconstruction error is
$$\sqrt{(-3 + 2.8)^2 + (2 - 1.4)^2} = \sqrt{0.4} = \frac{2}{\sqrt{10}} \tag{6}$$

4 PCA Generalizations

4.1 Robust PCA

Rubric: {code:10}

If you run `python main -q 4.1` the code will load a dataset X where each row contains the pixels from a single frame of a video of a highway. The demo applies PCA to this dataset and then uses this to reconstruct the original image. It then shows the following 3 images for each frame:

1. The original frame.
2. The reconstruction based on PCA.
3. A binary image showing locations where the reconstruction error is non-trivial.

Recently, latent-factor models have been proposed as a strategy for "background subtraction": trying to separate objects from their background. In this case, the background is the highway and the objects are the cars on the highway. In this demo, we see that PCA does an OK job of identifying the cars on the highway in that it does tend to identify the locations of cars. However, the results aren't great as it identifies quite a few irrelevant parts of the image as objects.

Robust PCA is a variation on PCA where we replace the L2-norm with the L1-norm,

$$f(Z, W) = \sum_{i=1}^{n}\sum_{j=1}^{d} |\langle w^j, z_i\rangle - x_{ij}|,$$

and it has recently been proposed as a more effective model for background subtraction. Complete the class *pca.RobustPCA*, that uses a smooth approximation to the absolute value to implement robust PCA. Briefly comment on the results.

Note: in its current state, *pca.RobustPCA* is just a copy of *pca.AlternativePCA*, which is why the two rows of images are identical.

Hint: most of the work has been done for you in the class *pca.AlternativePCA*. This work implements an alternating minimization approach to minimizing the (L2) PCA objective (without enforcing orthogonality). This gradient-based approach to PCA can be modified to use a smooth approximation of the L1-norm. Note that the log-sum-exp approximation to the absolute value may be hard to get working due to numerical issues, and a numerically-nicer approach is to use the "multi-quadric" approximation:

$$|\alpha| \approx \sqrt{\alpha^2 + \epsilon},$$

where ϵ controls the accuracy of the approximation (a typical value of ϵ is 0.0001).

Answer: Please refer to pca.py for details. From the results, we notice that by using robust PCA, the loss is smaller at each iteration. In the binary image obtained from AlternativePCA, several irrelevant parts are identified as objects. By using RobustPCA, these irrelevant parts are excluded in the binary image.

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)?

Answer:

1. L1 loss is more robust to the outliers. Therefore, this approach can be more suitable to exclude the objects from the background.
2. n is related with the number of video frames, larger n corresponds to larger number of video frames. d is related with the size of each frame. Larger d corresponds to larger size of each frame.
3. Smaller threshold leads to higher chance of false positives, meaning that many parts that are not cars are identified as cars. Larger threshold leads to higher chance of false negative, meaning that real cars may not be identified.

5 Very-Short Answer Questions

Rubric: {reasoning:11}

1. Assuming we want to use the original features (no change of basis) in a linear model, what is an advantage of the "other" normal equations over the original normal equations?
2. In class we argued that it's possible to make a kernel version of k-means clustering. What would an advantage of kernels be in this context?
3. In the language of loss functions and regularization, what is the difference between MLE and MAP?
4. What is the difference between a generative model and a discriminative model?
5. With PCA, is it possible for the loss to increase if k is increased? Briefly justify your answer.
6. What does "label switching" mean in the context of PCA?
7. Why doesn't it make sense to do PCA with $k > d$?
8. In terms of the matrices associated with PCA (X, W, Z, $\hat{X}$), where would an "eigenface" be stored?
9. What is an advantage and a disadvantage of using stochatic gradient over SVD when doing PCA?
10. Which of the following step-size sequences lead to convergence of stochastic gradient to a stationary point?
    (a) $\alpha^t = 1/t^2$.
    (b) $\alpha^t = 1/t$.
    (c) $\alpha^t = 1/\sqrt{t}$.
    (d) $\alpha^t = 1$.
11. We discussed "global" vs. "local" features for e-mail classification. What is an advantge of using global features, and what is advantage of using local features?

Answer:

1. Make it faster
2. Able to identify nonlinear structure
3. MLE is the maximum likelihood estimation $p(D|w)$ which is only related with the loss functions. MAP is Maximum a Posteriori $p(w|D)$, which is related with both the loss functions and regularization

4. The generative model learns the probability distribution $p(x, y)$, while discriminative model learns $p(y|x)$
5. No, increasing k will not increase the loss. Higher
6. Label switching is the reason why PCA is non-uniqueness
7. Take a picture as an example, d represents the total number of pixels of the picture, k represents the number of parts of the picture as basis. Therefore, the largest number of k should be the number of pixels. I doesn't make sense to do PCA with $k > d$ Moreover, we can say that when k is large, the data is not compressed and it is meaningless to
8. The eigenface is stored in W
9. Advantage: the computational complexity will be low especially when X is large. Disadvantage: There are hyperparameters needed to be optimized
10. (a) (b) (c)
11. Global features can predict for new users whose information is not known. Local feature is more suitable for customization.