CPSC 340 Assignment 5 (due Friday March 20 at 11:55pm)

Instructions

Rubric: {mechanics:5}

IMPORTANT!!! Before proceeding, please carefully read the general homework instructions at `https://www.cs.ubc.ca/~fwood/CS340/homework/`. The above 5 points are for following the submission instructions. You can ignore the words "mechanics", "reasoning", etc.

We use blue to highlight the deliverables that you must answer/do/submit with the assignment.

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).

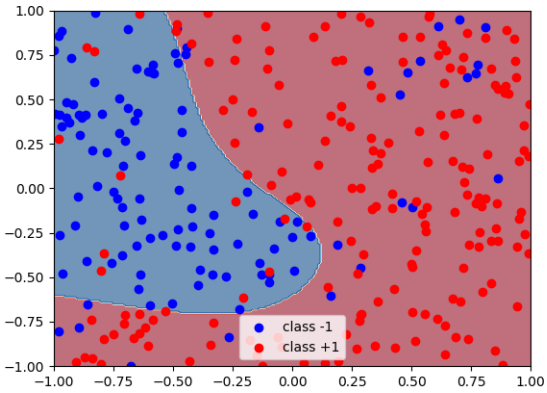
1.1 Implementing kernels

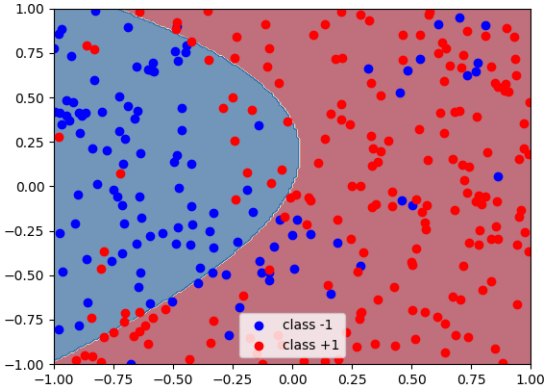
Rubric: {code:5}

Implement the polynomial kernel and the RBF kernel for logistic regression. Make sure to include the code you have written in your pdf GradeScope submission. 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: RBF Kernel - training error 0.127, validation error 0.090 and polynomial RBF - training error 0.183, validation error 0.170

```
def kernel_RBF(X1, X2, sigma=0.5):
    n,d = X1.shape
    N,D = X2.shape
    K = np.zeros((n,N))
    for i in range(n):
        for j in range(N):
            K[i, j] = np.exp(- (np.linalg.norm(X1[i] - X2[j])) ** 2 / (2 * sigma ** 2))
    return K

def kernel_poly(X1, X2, p=2):
    return (1+np.dot(X1,X2.T))**p
```





1.2 Hyperparameter search

Rubric: {code:3}

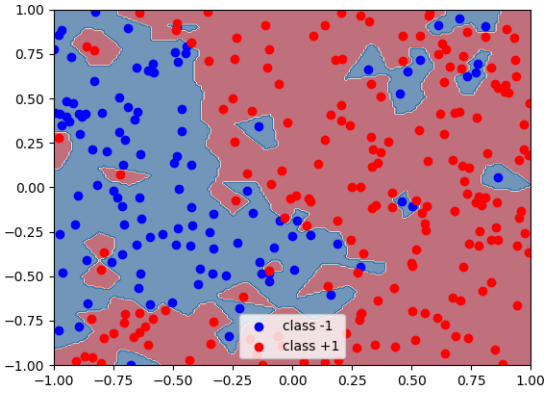
For the RBF kernel logistic regression, consider the hyperparameters values $\sigma = 10^m$ for $m = -2, -1, \ldots, 2$ and $\lambda = 10^m$ for $m = -4, -3, \ldots, 0$. In `main.py`, sweep over the possible combinations of these hyperparameter values. Make sure to include the code you have written in your pdf GradeScope submission. 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.

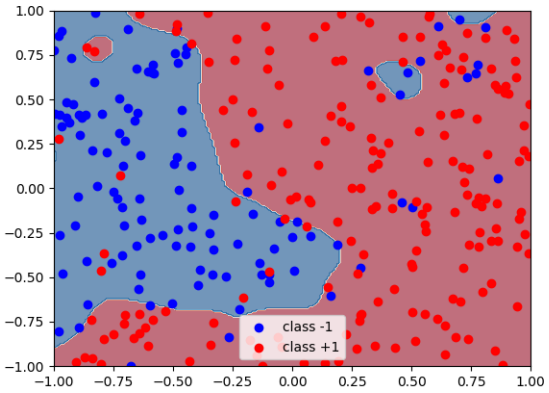
Note: on the job you might choose to use a tool like scikit-learn's `GridSearchCV` to implement the grid search, but here we are asking you to implement it yourself by looping over the hyperparameter values.

Answer: Best training error: $\sigma = 0.01, \lambda = 0.0001$. Best validation error: $\sigma = 0.1, \lambda = 1$

```
dataset = load_dataset('nonLinearData.pkl')
    X = dataset['X']
    y = dataset['y']

    Xtrain, Xtest, ytrain, ytest = train_test_split(X,y,random_state=0)

    sig = 10 ** np.arange(-2, 3, dtype=float)
    lam = 10 ** np.arange(-4, 1, dtype=float)
```

```
tr_err = np.inf
val_err = np.inf
tr_best = []
val_best = []

for s in sig:
    for l in lam:
        RBF_kernel = kernelLogRegL2(sigma = s,kernel_fun=kernel_RBF,lammy = l)
        RBF_kernel.fit(Xtrain,ytrain)
        loc_tr = np.mean(RBF_kernel.predict(Xtrain) != ytrain)
        loc_val = np.mean(RBF_kernel.predict(Xtest) != ytest)
        if loc_tr < tr_err:
            tr_err = loc_tr
            tr_best = [s, l]
        if loc_val < val_err:
            val_err = loc_val
            val_best = [s, l]

print("best training error parameters: ", tr_best)
print("training error", tr_err)
print("best validation error parameters: ", val_best)
print("validation error", val_err)

RBF_kernel = kernelLogRegL2(sigma=tr_best[0], kernel_fun=kernel_RBF, lammy=tr_best[1])
RBF_kernel.fit(Xtrain, ytrain)
utils.plotClassifier(RBF_kernel, Xtrain, ytrain)
utils.savefig("RBFbesttraining.png")

RBF_kernel = kernelLogRegL2(sigma=val_best[0], kernel_fun=kernel_RBF, lammy=val_best[1])
RBF_kernel.fit(Xtrain, ytrain)
utils.plotClassifier(RBF_kernel, Xtrain, ytrain)
utils.savefig("RBFbestvalidation.png")
```





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: Lambda with $m = -4$ and sigma with $m = -2$ yielded the optimal result for training error. This caused overfitting, which can be seen in the plot above. On the other hand, Lambda with $m = 0$ and sigma with $m = -1$ yielded the optimal result for validation error. In regularization, as λ increases, training error increases, but it reduces overfitting by decreasing the overall slopes, which is implied from our results with lower validation error. Also, as σ increases, the Gaussian distribution will have more variance, and vice versa. More variance would imply that we are generalizing the result so larger σ will reduce overfitting, and hence better validation error but lower training error. Our results match these assumptions.

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.

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

Answer: $-\sum \log(\frac{1}{2}\exp(-|w^T x_i - y_i|) = -\sum[\log(\frac{1}{2}) - |w^T x_i - y_i|] = \text{constant} + ||Xw - y||_1$
$-\log(p(w)) =$
$p(w) = \Pi p(w_j) \propto \Pi \frac{1}{\sqrt{2\pi}\sigma}\exp(\frac{-w_j^2}{2\sigma^2}) = (\frac{1}{\sqrt{2\pi}\sigma})^d \exp(\frac{-\sum w_j^2}{2\sigma^2})$
$-\log(p(w)) = -d\log(\frac{1}{\sqrt{2\pi}\sigma}) - \log(\exp(\frac{-||w||^2}{2\sigma^2})) = \text{constant} + \frac{||w||^2}{2\sigma^2}$
$f(w) = ||Xw - y||_1 + \frac{||w||^2}{2\sigma^2}$

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.
Answer: $-\sum \log(\frac{1}{\sqrt{2\sigma_i^2\pi}}\exp(-\frac{(w^T x_i - y_i)^2}{2\sigma_i^2})) = -\sum[\log(\frac{1}{\sqrt{2\sigma_i^2\pi}}) + \log(\exp(-\frac{(w^T x_i - y_i)^2}{2\sigma_i^2}))] = \text{constant} +$
$\frac{1}{2}\sum \frac{1}{\sigma_i^2}(w^T x_i - y_i)^2 = \text{constant} + \frac{1}{2}(Xw - y)^T \Sigma^{-1}(Xw - y)$
$-\log(p(w)) = -\log((\frac{\lambda}{2})^d) - \log(\exp(-\lambda||w||_1)) = \text{constant} + \lambda||w||_1$
$f(w) = \frac{1}{2}(Xw - y)^T \Sigma^{-1}(Xw - y) + \lambda||w||_1$

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).
Answer: $p(w) \propto \Pi\sqrt{\frac{\lambda}{2\pi}}\exp(-\lambda\frac{w_j^2}{w}) = \sqrt{\frac{\lambda}{2\pi}}^d \exp(-\frac{\lambda}{2}\sum w_j^2)$
$-\log(p(w)) = -[d\log(\sqrt{\frac{\lambda}{2\pi}}) - \frac{\lambda}{2}\sum w_j^2] = \text{constant} + \frac{\lambda}{2}||w||^2$
$-\sum \log(p(y_i|x_i, w)) = \text{constant} + \frac{\nu+1}{2}\sum \log(1 + \frac{(w^T x_i - y_i)^2}{2})$
$f(w) = \frac{\nu+1}{2}\sum \log(1 + \frac{(w^T x_i - y_i)^2}{2}) + \frac{\lambda}{2}||w||^2$

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: $-\sum \log(\frac{\exp(y_i w^T x_i)\exp(-\exp(w^T x_i))}{y!}) = -\sum[(y_i w^T x_i) - \exp(w^T x_i) - \log(y_i!)] = \sum[-y_i w^T x_i +$
$\exp(w^T x_i)] + \text{constant} = -yXw + \sum \exp(w^T x_i) + \text{constant}$
$f(w) = -yXw + \sum \exp(w^T x_i)$

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.

1. $\mu_1 = 0$, $\mu_2 = 1$
   $x_1 - \mu_1 = x_1$, $x_2 - \mu_2 = (2, 0, 1, -2, -1)$
   $x_1 = -2(x_2 - \mu_2)$

$$\tilde{X} = \begin{pmatrix} -4 & 2 \\ 0 & 0 \\ -2 & 1 \\ 4 & -2 \\ 2 & -1 \end{pmatrix}$$

   $(U, S, V) = svd(\tilde{X})$ where $V = \begin{bmatrix} 0.8944 & -0.4472 \\ -0.8944 & -0.4472 \end{bmatrix}$
   Therefore, $w = \begin{bmatrix} 0.8944, -0.4472 \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{5}}, -\frac{1}{\sqrt{5}} \end{bmatrix}$

2. The centered point is $\tilde{x} = (-3, 1.5)$
   $\tilde{Z} = \tilde{x}w^T(ww^T)^{-1} = -\frac{6}{\sqrt{5}} - \frac{1.5}{\sqrt{5}} = -\frac{7.5}{\sqrt{5}} = -\frac{3\sqrt{5}}{2}$
   Loss is $||\tilde{Z}w - \tilde{x}||^2 = || - \frac{3\sqrt{5}}{2}(\frac{2}{\sqrt{5}}, -\frac{1}{\sqrt{5}}) - (-3, 1.5)||^2 = 0$

3. The centered point is $\tilde{x} = (-3, 1)$
   $\tilde{Z} = \tilde{x}w^T(ww^T)^{-1} = -\frac{6}{\sqrt{5}} - \frac{1}{\sqrt{5}} = -\frac{7}{\sqrt{5}} = -\frac{7\sqrt{5}}{5}$
   Loss is $||\tilde{Z}w - \tilde{x}||^2 = ||(-\frac{14}{5}, \frac{7}{5}) - (-3, 1)||^2 = ||(\frac{1}{5}, \frac{2}{5})||^2 = \frac{1}{5}$

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,

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

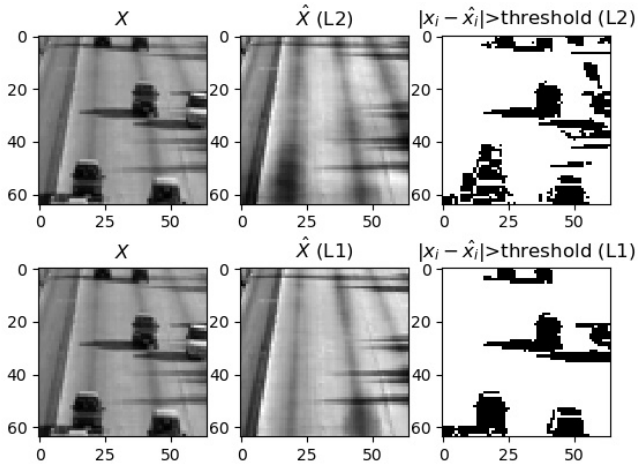
and it has recently been proposed as a more effective model for background subtraction. Complete the class *pca.RobustPCA*, that uses a smooth approximation to the absolute value to implement robust PCA. Make sure to include the code you have written in your pdf GradeScope submission. Briefly comment 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: Using a smooth approximation to implement a robust PCA decreases noise in the final results, evident from the differences between threshold(L2) and threshold(L1). In $\hat{X}$ of L2, there are wider and darker "shadows" with the cars removed. $\hat{X}$ of L1 is noticeably cleaner.

```
class RobustPCA(AlternativePCA):
    def _fun_obj_z(self, z, w, X, k):
        n, d = X.shape
        Z = z.reshape(n, k)
        W = w.reshape(k, d)
        R = np.dot(Z, W) - X

        alpha = np.sqrt(R ** 2 + 0.0001)
        f = np.sum(alpha)
        g = np.dot(R / alpha, W.transpose())
        return f, g.flatten()

    def _fun_obj_w(self, w, z, X, k):
        n, d = X.shape
        Z = z.reshape(n, k)
        W = w.reshape(k, d)
        eps = 0.0001
        R = np.dot(Z, W) - X

        alpha = np.sqrt(R ** 2 + 0.0001)
        f = np.sum(alpha)
        g = np.dot(Z.transpose(), R / alpha)
        return f, g.flatten()
```

4.2 Reflection

Rubric: {reasoning:3}

1. Briefly explain why using the L1 loss might be more suitable for this task than L2.
   Answer: L1 loss is more sensitive to outliers, which in this case are the cars. Hence, using L1 loss gives us a clearer distinction between the background (road) and the outliers (cars).

2. How does the number of video frames and the size of each frame relate to n, d, and/or k?
   Answer: n represents the number of video frames, and increases as we increase the number of frames. d represents the size of each frame in pixels (height*width), or the number of features. Hence, d increases as the size of the frame increases. k is a hyperparameter that is chosen by us.

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: The number of false positives increases as the threshold is decreased, or "relaxed" - more of the errors are considered significant "outliers". For the opposite reason, the number of false negatives increases as the threshold increases.

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?
   Answer: The "other" normal equations replace Z and $\tilde{Z}$ with K and $\tilde{K}$ to compute dot-products without the features, significantly reducing runtime: $O(n^2d + n^3)$ instead of $O(n^2d^p + n^3)$

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?
Answer: Applying the kernel trick for k-means clustering allows for non-convex clusters.

3. In the language of loss functions and regularization, what is the difference between MLE and MAP?
Answer: MAP estimation maximizes the probability of parameters 'w' given our data. MLE estimation maximizes the likelihood of our data given 'w'. MLEs are probabilistic interpretations of loss functions, while MAPs are probabilistic interpretation of regularizers.

4. What is the difference between a generative model and a discriminative model?
Answer: A generative model optimizes $p(y, X|w)$, which models 'X'. A discriminative model maximizes $p(y|X, w)$, while is conditioned on the fact that the features 'X' are fixed.

5. With PCA, is it possible for the loss to increase if k is increased? Briefly justify your answer.
Answer: No. As k increases, the variance of x_{ij} decreases. Hence, loss must always decrease if k increases.

6. What does "label switching" mean in the context of PCA?
Answer: The label-switching problem in PCA is due to the non-uniqueness of PCA. For example, if $k \geq 2$, $\text{span}(w_1, w_2) = \text{span}(w_2, w_1)$.

7. Why doesn't it make sense to do PCA with $k > d$?
Answer: k represents the number of "parts", while d represents the number of "features", which cannot be split further. It does not make sense to have more parts than there are number of features, so k is always $\leq d$.

8. In terms of the matrices associated with PCA (X, W, Z, $\hat{X}$), where would an "eigenface" be stored?
Answer: Each "eigenface" is stored in one row of X.

9. What is an advantage and a disadvantage of using stochastic gradient over SVD when doing PCA?
Answer: Stochastic gradient is computationally faster than computing SVD when doing PCA, but SVD enforces orthogonality for W.

10. Which of the following step-size sequences lead to convergence of stochastic gradient to a stationary point?

    (a) $\alpha^t = 1/t^2$. Answer: does not converge $\frac{\sum(\alpha^t)^2}{\sum \alpha^t} = \frac{\pi^2}{15} \neq 0$

    (b) $\alpha^t = 1/t$. Answer: converges $\frac{\sum(\alpha^t)^2}{\sum \alpha^t} = \frac{\pi^2/6}{\infty} = 0$

    (c) $\alpha^t = 1/\sqrt{t}$. Answer: converges $\frac{\sum(\alpha^t)^2}{\sum \alpha^t} = \frac{O(\sqrt{k})}{O(\log(k))} = 0$

    (d) $\alpha^t = 1$. Answer: does not converge $\frac{\sum(\alpha^t)^2}{\sum \alpha^t} = 1$

11. We discussed "global" vs. "local" features for e-mail classification. What is an advantage of using global features, and what is advantage of using local features?
Answer: Global features are features applicable to all "users". For example, these are useful for capturing "globally" important messages like "this is your mother, something terrible happened, call me ASAP." Local features are features for specific users. For example, not all users may consider a "pizza discount email" to be spam.