# Problem Set 5

CS 4375

Due: 12/09/2024 by 11:59pm

Note: all answers should be accompanied by explanations for full credit. Late homeworks will not be accepted. As always, no ML library routines are permitted.

## Problem 1: Logistic Regression (40pts)

For this problem, consider the Sonar data set attached to the problem set (the last column in the data is the class label).

1. Fit a logistic regression classifier to the training data set. What is the accuracy on the test set?
2. Fit a logistic regression classifier with an $\ell_2$ penalty on the weights to this data set using the validation set to select a good choice of the regularization constant. Report your selected constant, the learned weights and bias, and the accuracy on the test set.
3. Fit a logistic regression classifier with an $\ell_1$ penalty on the weights to this data set using the validation set to select a good choice of the regularization constant. Report your selected constant, the learned weights and bias, and the accuracy on the test set.
4. Does $\ell_1$ or $\ell_2$ tend to produce sparser weight vectors?
5. Generate your own linearly separable test data in $\mathbb{R}^2$ and fit a standard SVM model to the data.
   (a) Using your data, explain why in standard logistic regression, without any type of regularization, the weights may not converge (even though the predicted label for each data point effectively does) if the input data is linearly separable. (Hint: try it and see what happens)
   (b) Fit a logistic regression classifier with an $\ell_2$ penalty on the weights to this data set using the validation set to select a good choice of the regularization constant.
   (c) Plot your data, the SVM solution, and the logistic regression solution. Which model do you prefer for your data set?

## Problem 2: Gaussian Naïve Bayes (30pts)

For this problem, consider the Sonar data set attached to the problem set with features $x$ and label $y$. Suppose that you want to fit a Gaussian NB model to this data. That is, assume that you would

like to fit a probability distribution of the form

$$p(x_1, \ldots, x_n, y) = p(y) \prod_{i=1}^{n} p(x_i|y),$$

where $p(x_i|y)$ is a distinctly parameterized normal distribution, i.e., $p(x_i|y)$ is parameterized by $\mu_{i,y}$ and $\sigma_{i,y}$.

1. Given a data set with $m$ continuous features, what is the log-likelihood of the Gaussian NB model? Compute the MLE for each of the model parameters.
2. Fit a Gaussian NB model to the training data. What is the accuracy of your trained model on the test set?
3. What kind of prior might make sense for this model? Explain.
4. How does naïve Bayes compare to logisitc regression for this data? Which would you prefer?

## Problem 3: Gaussian Mixtures vs. $k$-means (30pts)

For this problem, use the `leaf.data` file provided with Homework 4. Again, the class labels are still in the data set and should be used for evaluation only (i.e., don't use them in the clustering procedure), but the specimen number has been removed. You should preprocess the data so that the non-label attributes have mean zero and variance one.

It is possible that during the Gaussian mixture updates, some of the iterated covariances matrcies might not be strictly positive definite, i.e., some of the eigenvalues might be too close to zero. This is problematic as zero variances cannot be allowed. To fix this, let's insist that all of the produced covariance matrices must have a minimum eigenvalue of $\epsilon > 0$. To do this, compute the eigendecomposition of each covariance matrix, say $QDQ^T$. If any of the eigenvalues on the diagonal of $D$ fall below $\epsilon$, then set them to $\epsilon$ by changing the corresponding element of the diagonal of $D$. This yields a new diagonal matrix, call it $F$. You can now set the covariance matrix to $QFQ^T$.

Train a Gaussian mixture model for each $k \in \{10, 20, 30, 36\}$ starting from twenty different random initializations (use the $k$-means++ strategy to select the starting means and set the starting covariance matrices equal to the identity matrix) for each $k$. Report the mean and variance of the converged log-likelihood for each $k$.