EE 511 – Assignment 2

Due: 26 Jan '19

Part I: written Assignment (6 points)

1. Assume that random variable Y has a probability distribution $p(y|x) \sim N(w^t x, \sigma^2)$, where x and w are d-dimensional.

   (a) Find the maximum likelihood estimate of w given data $\{(x_i, y_i); i = 1, \ldots, n\}$.

   (b) Find the MAP estimate of w assuming a Gaussian prior for W that is zero mean and has covariance $\sigma_0 I$. Explain how this relates to ridge regression.

2. Feature normalization (subtract the feature mean and divide by the standard deviation) is typically used when using regularization. Explain why feature normalization is important for regularization but not for ordinary least squares regression.

3. Each set of class-conditional distributions below correspond to a different classification problem, where x is the feature used for classification c_j is the class label. For each problem, determine whether the distribution assumption corresponds to a Naive Bayes classifer and whether it can be implemented with a linear decision function. There are m classes, and $p(Y = c_j) = q_j$.

   (a) $x \in \Re^d$ and $p(x|c_j) \sim N(\mu_j, \sigma_j^2 I)$

   (b) $x \in \Re^2$ and $p(x|c_j) \sim N(\mu_j, \Sigma)$, where $\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$

   (c) $x_1 \in \Re,\ x_2 \in \{0, 1\}$, $p(x_1, x_2|c_j) = p(x_1|c_j)p(x_2|c_j)$, and

   $$p(x_1|c_j) \sim N(\mu_j, \sigma^2) \quad p(x_2|c_j) \sim \text{Bernoulli with param } \alpha_j$$

   (d) $x_1 \in \Re,\ x_2 \in \Re^+$, $p(x_1, x_2|c_j) = p(x_1|c_j)p(x_2|c_j)$, and

   $$p(x_1|c_j) = \frac{\lambda_j}{2} e^{-\lambda_j |x_1|} \quad p(x_2|c_j) = \alpha_j e^{-\alpha_j x_2}$$

Part II: Computer Assignment (14 points)

1. Load the data from the AmesHousing.txt file. There should be 2931 rows.

   The file can be downloaded from `https://ww2.amstat.org/publications/jse/v19n3/decock/AmesHousing.txt`. A description of the data is available at `https://ww2.amstat.org/publications/jse/v19n3/Decock/DataDocumentation.txt`.

2. Preprocessing:

   There are some missing values in this data. Replace all the missing values for numerical features with zeros and for categorical features use a special string to indicate a missing value.

   The numerical features are:

   ```
   numerical_variables = [’Lot Area’, ’Lot Frontage’, ’Year Built’,
                          ’Mas Vnr Area’, ’BsmtFin SF 1’, ’BsmtFin SF 2’,
                          ’Bsmt Unf SF’, ’Total Bsmt SF’, ’1st Flr SF’,
                          ’2nd Flr SF’, ’Low Qual Fin SF’, ’Gr Liv Area’,
                          ’Garage Area’, ’Wood Deck SF’, ’Open Porch SF’,
                          ’Enclosed Porch’, ’3Ssn Porch’, ’Screen Porch’,
                          ’Pool Area’]
   ```

   The categorical features are:

   ```
   discrete_variables = [’MS SubClass’, ’MS Zoning’, ’Street’,
                         ’Alley’, ’Lot Shape’, ’Land Contour’,
                         ’Utilities’, ’Lot Config’, ’Land Slope’,
                         ’Neighborhood’, ’Condition 1’, ’Condition 2’,
                         ’Bldg Type’, ’House Style’, ’Overall Qual’,
                         ’Overall Cond’, ’Roof Style’, ’Roof Matl’,
                         ’Exterior 1st’, ’Exterior 2nd’, ’Mas Vnr Type’,
                         ’Exter Qual’, ’Exter Cond’, ’Foundation’,
                         ’Bsmt Qual’, ’Bsmt Cond’, ’Bsmt Exposure’,
                         ’BsmtFin Type 1’, ’Heating’, ’Heating QC’,
                         ’Central Air’, ’Electrical’, ’Bsmt Full Bath’,
                         ’Bsmt Half Bath’, ’Full Bath’, ’Half Bath’,
                         ’Bedroom AbvGr’, ’Kitchen AbvGr’, ’Kitchen Qual’,
                         ’TotRms AbvGrd’, ’Functional’, ’Fireplaces’,
                         ’Fireplace Qu’, ’Garage Type’, ’Garage Cars’,
                         ’Garage Qual’, ’Garage Cond’, ’Paved Drive’,
                         ’Pool QC’, ’Fence’, ’Sale Type’, ’Sale Condition’]
   ```

3. Split the data into train, validation and test sets. We will do this by using the Order column in the data file. For the validation set take the examples where the Order mod 5 is 3 and for the test set use the examples where the Order mod 5 is 4. The rest is for training.

4. Now let's do a simple one variable least squares linear regression as a warm-up. Predict the sale price based on the "Gr Liv Area" feature. Make a scatter plot of this feature vs. the sale price using the training data and overlay the line from your model. What is the equation for the line that you found?

   Apply the model to the data from the validation set and compute the root mean sequared error (RMSE).

5. Now that we have our simple model working, let's add more features.

   First, transform the categorical features to a one-hot encoding so that they can be used in the model. For example, the "Alley" column can take on three possible values "Pave", "Grvl", and "Missing". This will become a 3-dimensional one-hot vector.

   Once the categorical features have been transformed, concatenate them with the numerical features and train a new model. Compute the RMSE on the validation set for this model.

6. We can improve the model by using L1 regularization, i.e. penalizing the absolute value of the coefficients. When L1 regularization is used for linear regression it is called Lasso Regression. We need to use cross-validation to select the value of alpha, which is the weight on the L1 regularization term.

   The first thing to do here is to normalize the features by subtracting the mean and dividing by the standard deviation. Be sure to use the mean and variance estimated from the training data when normalizing the validation (and test) data.

   Using cross validation, train models for values of alpha ranging from 50 to 500 in intervals of 50. Make a plot of the RMSE on the train and vaidation sets for each value of alpha. Which setting has the lowest error on the validation set? Briefly explain the concept of over-fitting and how this graph can be used to detect it.

   Plot the number of non-zero coefficients in the model for each of the values of alpha that you tried.

7. Now it's time to use the test data. Take the single variable model you trained, the least squares model that uses all of the variables, and the regularized model that you selected and apply them to the test data. Make a table and report the RMSE for each condition for both the validation set and the test set.