# Machine Learning Computer Lab 1 (Group A7)

Qinyuan Qi(qinqi464) Satya Sai Naga Jaya Koushik Pilla (satpi345)
Daniele Bozzoli(danbo826)

2023-11-07

## Assignment 1: Handwritten digit recognition with K- nearest neighbors.

The data file optdigits.csv contains information about normalized bitmaps of handwritten digits from a preprinted form from a total of 43 people. The data were first derived as 32x32 bitmaps which were then divided into nonoverlapping blocks of 4x4 and the number of on pixels are counted in each block. This has generated the resulting image of size 8x8 where each element is an integer in the range 0..16. Accordingly, each row in the data file is a sequence corresponding to 8x8 matrix, and the last element shows the actual digit from 0 to 9.

1. Import the data into R and divide it into training, validation and test sets (50%/25%/25%) by using the partitioning principle specified in the lecture slides.

2. Use training data to fit 30-nearest neighbor classifier with function *kknn*() and kernel="rectangular" from package *kknn* and estimate

- Confusion matrices for the training and test data (use table())
- Misclassification errors for the training and test data Comment on the quality of predictions for different digits and on the overall prediction quality.

3. Find any 2 cases of digit "8" in the training data which were easiest to classify and 3 cases that were hardest to classify (i.e. having highest and lowest probabilities of the correct class). Reshape features for each of these cases as matrix 8x8 and visualize the corresponding digits (by using e.g. heatmap() function with parameters Colv=NA and Rowv=NA) and comment on whether these cases seem to be hard or easy to recognize visually.

4. Fit a K-nearest neighbor classifiers to the training data for different values of $k = 1, 2, ..., 30$ and plot the dependence of the training and validation misclassification errors on the value of K (in the same plot). How does the model complexity change when $K$ increases and how does it affect the training and validation errors? Report the optimal $K$ according to this plot. Finally, estimate the test error for the model having the optimal $K$, compare it with the training and validation errors and make necessary conclusions about the model quality.

5. Fit K-nearest neighbor classifiers to the training data for different values of $K = 1, 2, ..., 30$, compute the error for the validation data as cross-entropy (when computing log of probabilities add a small constant within log, e.g. 1e-15, to avoid numerical problems) and plot the dependence of the validation error on the value of $K$. What is the optimal $K$ value here? Assuming that response has multinomial distribution, why might the cross-entropy be a more suitable choice of the error function than the misclassification error for this problem?

**Answer:**

## Assignment 2: Linear regression and ridge regression

The data file parkinson.csv is composed of a range of biomedical voice measurements from 42 people with early-stage Parkinson's disease recruited to a six-month trial of a telemonitoring device for remote symptom progression monitoring. The purpose is to predict Parkinson's disease symptom score (motor UPDRS) from the following voice characteristics: - Jitter(%),Jitter(Abs),Jitter:RAP,Jitter:PPQ5,Jitter:DDP - Several measures of variation in fundamental frequency - Shimmer,Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,Shimmer:APQ11,Shimmer:DDA - Several measures of variation in amplitude - NHR,HNR - Two measures of ratio of noise to tonal components in the voice - RPDE - A nonlinear dynamical complexity measure - DFA - Signal fractal scaling exponent - PPE - A nonlinear measure of fundamental frequency variation

1. Divide it into training and test data (60/40) and scale it appropriately. In the coming steps, assume that motor_UPDRS is normally distributed and is a function of the voice characteristics, and since the data are scaled, no intercept is needed in the modelling.
2. Compute a linear regression model from the training data, estimate training and test MSE and comment on which variables contribute significantly to the model.
3. Implement 4 following functions by using basic R commands only (no external packages):

a. *Logklilihood* function that for a given parameter vector $\theta$ and dispersion $\sigma$ computes the log-likelihood function $logP(T|\theta,\sigma)$ for the stated model and the training data.

b. *Ridge* function that for given vector $\theta$ , scalar $\sigma$ and scalar $\lambda$ uses function from 3a and adds a Ridge penalty $\lambda||\theta||^2$ to the minus log-likelihood.

c. *RidgeOpt* function that depends on scalar $\lambda$, uses function from 3b and function optim() with method="BFGS" to find the optimal $\theta$ and $\sigma$ for the given $\lambda$.

d. *DF* function that for a given scalar $\lambda$ computes the degree of freedom of the Ridge model based on the training data.

4. By using function *RidgeOpt*, compute optimal $\theta$ parameters for $\lambda = 1$,$\lambda = 100$,$\lambda = 1000$. Use the estimated parameters to predict the motor_UPDRS values for training and test data and report the training and test MSE values. Which penalty parameter is most appropriate among the selected ones? Compute and compare the degrees of freedom of these models and make appropriate conclusions.

**Answer:**

## Assignment 3. Logistic regression and basis function expansion

The data file pima-indians-diabetes.csv contains information about the onset of diabetes within 5 years in Pima Indians given medical details. The variables are (in the same order as in the dataset): 1. Number of times pregnant. 2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test. 3. Diastolic blood pressure (mm Hg). 4. Triceps skinfold thickness (mm). 5. 2-Hour serum insulin (mu U/ml). 6. Body mass index (weight in kg/(height in m)^2). 7. Diabetes pedigree function. 8. Age (years). 9. Diabetes (0=no or 1=yes).

1. Make a scatterplot showing a Plasma glucose concentration on Age where observations are colored by Diabetes levels. Do you think that Diabetes is easy to classify by a standard logistic regression model that uses these two variables as features? Motivate your answer.
2. Train a logistic regression model with $y$ = Diabetes as target $x_1$ = Plasma glucose concentration and $x_2$ = Age as features and make a prediction for all observations by using $r = 0.5$ as the classification threshold. Report the probabilistic equation of the estimated model (i.e., how the target depends on the features and the estimated model parameters probabilistically). Compute also the training misclassification error and make a scatter plot of the same kind as in step 1 but showing the predicted values of Diabetes as a color instead. Comment on the quality of the classification by using these results.
3. Use the model estimated in step 2 to a) report the equation of the decision boundary between the two classes b) add a curve showing this boundary to the scatter plot in step 2. Comment whether the decision boundary seems to catch the data distribution well.
4. Make same kind of plots as in step 2 but use thresholds $r = 0.2$ and $r = 0.8$. By using these plots, comment on what happens with the prediction when $r$ value changes.

5. Perform a basis function expansion trick by computing new features $z_1 = x_1^4$,$z_2 = x_1^3 x_2$,$z_3 = x_1^2 x_2^2$,$z_4 = x_1 x_2^3$,$z_5 = x_2^4$ adding them to the data set and then computing a logistic regression model with $y$ as target and $x_1, x_2, x_3, x_4, x_5$ as features. Create a scatterplot of the same kind as in step 2 for this model and compute the training misclassification rate. What can you say about the quality of this model compared to the previous logistic regression model? How have the basis expansion trick affected the shape of the decision boundary and the prediction accuracy?

**Answer:**

# Appendix: All code for this report

```
knitr::opts_chunk$set(echo = TRUE)
```