Homework set #1 - K nearest neighbors

Ido B. Gattegno, Haim H. Permuter

August 31, 2016

1 Probability and Markov chains

Three random variables X, Y, Z form a Markov chanin $X \to Y \to Z$ if

$$P(x, y, z) = P(x)P(y|x)P(z|y). \tag{1}$$

Similarly, $X \leftarrow Y \leftarrow Z$ if

$$P(x, y, z) = P(z)P(y|z)P(y|x). \tag{2}$$

1. **Markov chain def.** Show that $X \to Y \to Z$ if and only if $X \leftarrow Y \leftarrow Z$. Because of this property we denote a Markov chain simply as $X - Y - Z$.
2. **True or False.** If its true provide a proof if its false provide a counter example.
   (a) Let X, Y, Z, W be r.v. that satisfies the Markov $X - Y - Z - W$. Then $X - Y - Z$ and $Y - Z - W$ hold.
   (b) Let X, Y, Z, W be r.v. that satisfies the Markov $X - Y - Z$ and $Y - Z - W$. Then $X - Y - Z - W$ hold.
   (c) Let $P(x, y, z, w) = P(x)p(y|x)P(z, w)$ for all $x, y, z, w \in \mathcal{X} \times Y \times Z \times W$. Then $X - Y - (Z, W)$.
   (d) $X \perp\!\!\!\perp Y$ and $Y \perp\!\!\!\perp Z$ implies $X \perp\!\!\!\perp Z$.
   (e) If one can show that $P(x|y, z)$ is a function of only (x, y) then $X - Y - Z$ holds.
3. **Conditional Independence.** Let (X_i, θ_i) where $i = 1, 2, ..., n$ be a i.i.d distributed according to $P(x, \theta)$ where x is a feature and θ is its class. In addition let the pair $(X, \theta) \sim P(x, \theta)$ and independent of (X^n, θ^n) where X^n denotes $(X_1, X_2, ..., X_n)$ and θ^n denotes $(\theta_1, \theta_2, ..., \theta_n)$. Let x' be the element in x^n that is closest to x and θ' the associated class.
   (a) Is $X_2 \perp\!\!\!\perp X$?
   (b) Is $X' \perp\!\!\!\perp X$?
   (c) Is $\theta_1 \perp\!\!\!\perp \theta$?
   (d) Is $\theta' \perp\!\!\!\perp \theta$?
   (e) Is $\theta - X - X^n$?
   (f) Is $\theta - X - (X^n, X')$?
   (g) Is $\theta - X - (X^n, X', \theta^n)$?
   (h) Is $\theta - X - (X', \theta')$?
   (i) Prove that
   $$P(\theta, \theta'|x, x') = P(\theta|x)P(\theta'|x')$$
4. **Random variable with index** Let $A_1 - B_1 - C$ and $A_2 - B_2 - C$ and in addition $P_{A_1,B_1}(a, b) = P_{A_2,B_2}(a, b)$ for all a, b in the alphabet. Let $i(C)$ be a binary (deterministic) function of C that emits 1 or 2.
   (a) Show that $A_{i(C)} - B_{i(C)} - C$ holds.
   (b) Does it also hold if $P_{A_1,B_1}(a, b) \neq P_{A_2,B_2}(a, b)$.

2 K-Nearest Neighbors - algorithm and synthetic data

The *K-Nearest Neighbors* algorithm is a simple classification method, which uses a distance metric in order to determine the k samples that are closest to the test sample.

Assume that $\mathcal{T} \triangleq \{(x(i), l(i)),\ i = 1, \ldots, n\}$ is a training set of samples $x(i)$ and corresponding labels $l(i)$. Let $d(x, \tilde{x})$ denote a metric measure between two samples, x and $\tilde{x}$.
The nearest neighbor of $\tilde{x}$ in $\mathcal{T}$ with respect to $d(\cdot, \cdot)$ is

$$x_{1-nn} = \arg \min_{x':\ (x', l') \in \mathcal{T}} d(\tilde{x}, x')$$

The second nearest neighbor is the closest in $\mathcal{T}$ without x_{1-nn} and so on.

A common distance metric is the euclidean, i.e., $d(x, \tilde{x}) = ||x - \tilde{x}||_2$.
Once having k nearest neighbors, a decision method is applied to determine which class to predict.
For instance, the predicted class can be taken by majority.

1. Let

   $$\mu_1 = [0, -1, 1]^\top,\ \Sigma_1 = \mathbb{I}_{3\times 3}$$
   $$\mu_2 = [0, -3, 3]^\top,\ \Sigma_2 = \mathbb{I}_{3\times 3}$$

   where $\mathbb{I}$ is the identity matrix.
   Create $60,000$ synthetic samples from 2 Gaussian distribution, according the above parameters, each one belongs to a class.
   Namely, observations from class 1 are distributed according to $\mathcal{N}(\mu_1, \Sigma_1)$, and observations from class 2 according to $\mathcal{N}(\mu_2, \Sigma_2)$. Assume $P_L(1) = P_L(2) = 0.5$.

2. Plot 2 out of the three coordinates of all samples. Each class should appear in a different color.

3. Generate $10,000$ synthetic samples for test.

4. Apply K-NN algorithm with $k = 1$ on the test samples.
   We define classification error rate by the average loss, i.e.,

   $$p_e \triangleq \frac{\sum_{i=1}^{n} \mathbb{1}\left\{\hat{l}(\mathbf{x}(i)) \neq l(\mathbf{x}(i))\right\}}{n}$$

   where n is the number of test samples, $\hat{l}(\mathbf{x}(i))$ is the classification of the i-th observation from the test, and $l(\mathbf{x}(\mathbf{i}))$ is the true class of that observation.
   What is the error rate of your classification? Compare it with the error rate of Maximum Likelihood Estimator (MLE) with optimal threshold.

5. Repeat last step for $k = 2, 3, \ldots, 10$ and plot the error rate p_e against k.
   Is the error rate decreases with k?
   Should the error rate *always* decrease with k?

6. Let m be the number of training samples. Fix k and n, and plot the error rate as a function of m. How do you expect the graph to behave?

7. Replace the distance metric with $||\mathbf{x} - \tilde{\mathbf{x}}||_p$, for $p = 1, 2, \ldots, 10$ and plot the error rate against p.
   Does the distance metric effect the error rate?

3 MNIST database

We are motivated to classify handwritten digits, using *machine learning* algorithms. A well known benchmark database is the MNIST. It contains a large number of handwritten digits, centered and normalized into a fixed-size images (28×28 pixels).
Please download the example MATLAB files from:
`http://www.ee.bgu.ac.il/~idobenja/MachineLearning/Databases/MNIST.zip`.

Run `example.m` to load the MNIST database and plot some of the images.
At your workspace, the following variables will appear:

- `database_train_images` - 60000×784 matrix. Each row is an image, of size 28×28, unrolled into a vector.
- `database_train_labels` - 60000×1 vector. Each row is the digit that corresponds to `database_train_images`.
- `database_test_images` - 10000×784 matrix with test images.
- `database_test_labels` - 10000×1 vector with corresponding digits.

3.1 Visualization of data from MNIST database

Many machine learning problems come from real-life applications. If possible, it is important to visualize the data, in order to gain intuition, and learn what methods we want to apply. In our case, the data observations are 28×28 gray-scale images of digits. The provided example also provides a function `show_numbers.m`. This function plots an image from the data vector.

1. Plot the first 25 images in the training set. Does all images of 9 look alike?
2. Plot 10 random samples from the test, without revealing the corresponding digits from the label, and try to guess them. Were all of your guesses correct?

Experience the data some more before moving forward. Find 2 different digits that look alike. Find 2 samples of the same digits, that don't look alike at all.

3.2 Classification using KNN on MNIST database

1. What is the nearest neighbor of a train sample, assuming it is included in the training set?
2. Write code in MATLAB for k-nearest neighbors with euclidean metric. Find 5 nearest neighbors for the first 10 test samples and plot them together.
3. Test all 10000 digits in the test database with k nearest neighbors. Take decisions by majority. What is the error rate for $k = 5$? try various options for k.
4. What is the confusion matrix of the best test so far?
5. Since the computation time depends on the size of the train set, one suggested to take only 10000 train samples, instead of 60000. Plot the error rate against the size of train set.

We aim to further improve our previous results of classification.

1. Plot the k nearest neighbors of some of misclassified samples.
2. Replace the euclidean metric with various other metrics and test them. Summarize the results in a table.
3. Replace the majority decision with your own method. What are the results? try to combine with various distance metrics and k.
4. What is the lowest error rate you achieved in this exercise?