Stony Brook University
CSE512 – Machine Learning – Spring 17
Homework 1, Due: Feb, 21, 2017, 11:59PM

Instructions

- The homework is due on February 21, 2017. Anything that is received after the deadline will not be considered.
- The write-up must be prepared in Latex, including the Matlab code and figures in the report.
- If the question requires you to implement a Matlab function, submit the code and make sure it is sufficiently well documented that the TAs can understand what is happening.
- Each Question, regardless of how many sub-questions contains, is worth 10 points.

1 Linear algebra

1.1 Question 1

- Prove that $A \in \mathbb{R}^{n\times n}$ and $A^\top$ have the same eigenvalues.
- Let λ_i are the eigenvalues of $M \in \mathbb{R}^{n\times n}$. Determine the eigenvalues of $\alpha M + \beta I$, where I is the identity matrix, and $\alpha, \beta \in \mathbb{R}$.

2 Basic Statistics

This will help you get better familiarity with probability distributions.

2.1 Question 2: Probabilities

Denote by A and B events, and by $\bar{A}$ the complement of A. Prove the following:

- $\mathbb{P}(B \cap \bar{A}) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- If $A \subset B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$

2.2 Question 3: Gaussian Distribution

The Gaussian distribution of parameters μ and σ^2 is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \ .$$

Prove that the mean and the variance are μ and σ^2 respectively.

2.3 Question 4: Poisson Distribution

The Poisson distribution is used to model a flow of events given only the average rate at which an event occurs. Examples of its use include the following:

- The number of incoming requests to a server.
- The number of mutations on a strand of DNA.
- The number of cars arriving at a traffic light.
- The number of raindrops in a given area in a given time interval.

The Poisson distribution has one parameter, the average rate $\lambda > 0$ and has probability mass function

$$f(k;) = Pr(X = x) = \frac{\lambda^k e^{-\lambda}}{k!} \ .$$

- Compute mean and variance of the Poisson distribution (hint: Taylor expansion for e^{λ}).
- Let $X_1 \sim Poisson(\lambda_1)$ and $X_2 \sim Poisson(\lambda_2)$ be independent random variables. Show that the random variable $Z = X_1 + X_2$ is Poisson-distributed and compute its mean.

2.4 Question 5: Estimators

Let $X_1, \cdots, X_n$ be i.i.d. random variables with mean μ and variance σ^2.

- Prove that

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$S_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2,$$

  are unbiased estimators of the mean and variance, that is, $\mathbb{E}[M_n] = \mu$ and $\mathbb{E}[S_n] = \sigma^2$.

- For a distribution of your choice (Gaussian, uniform, etc.) implement these estimators and empirically show on a plot that these estimators indeed converge to the true quantities μ and σ^2 as we increase the sample size n. Based on your plots guess how fast (in terms of powers of n) they converge to the correct quantities.

3 (Agnostic) PAC Learning

3.1 Question 6

Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \{0, 1\}$, and let $S = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_m, y_m)\}$ be a random sample from $\mathcal{D}$. Let

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{D}}[h(x) \neq y]$$
$$L_S(h) = \frac{1}{m}|\{i : h(\boldsymbol{x}_i) \neq y_i\}|$$

Let h_S and h^* be the hypotheses in $\mathcal{F}$ with minimum training and generalization error, respectively:

$$h_S = \arg\min_{h\in\mathcal{F}} L_S(h),$$
$$h^* = \arg\min_{h\in\mathcal{F}} L_{\mathcal{D}}(h) \ .$$

Be sure to keep in mind that, unlike h^*, h_S is a random variable that depends on the random sample S.

- Prove that
$$\mathbb{E}[L_S(h_S)] \leq L_{\mathcal{D}}(h^*) \leq \mathbb{E}[L_{\mathcal{D}}(h_S)] \ .$$

- We now make use of a stronger concentration inequality than the one used in class.

  Theorem 1 (McDiarmids inequality)**.** *Let a function f for which*
  $$\forall x_1, \cdots, x_m, x_i' : |f(x_1, \cdots, x_i, \cdots, x_m) - f(x_1, \cdots, x_i', \cdots, x_m)| \leq c_i,$$
  and $X_1, \cdots, X_m$ independent but not necessarily identically distributed random variables. Then the following holds
  $$\mathbb{P}\left[|f(X_1, \cdots, X_m) - \mathbb{E}\left[f(X_1, \cdots, X_m)\right]| \geq \epsilon\right] \leq 2 \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right) \ .$$

  Using this theorem prove that, with probability at least $1 - \delta$
  $$|L_S(f_S) - \mathbb{E}[L_S(f_S)]| \leq O\left(\sqrt{\frac{\ln(1/\delta)}{m}}\right),$$
  where there the constants hidden in the big-O notation do not depend on $|\mathcal{F}|$.

- Explain in words the meaning of what you proved and how we would expect training error to compare to test error when using a machine learning algorithm on actual data.

3.2 Question 7

Let $\mathcal{F}$ a hypothesis class of binary classifiers. Show that if $\mathcal{F}$ is agnostic PAC learnable, then $\mathcal{F}$ is PAC learnable as well. Furthermore, if A is a successful agnostic PAC learner for $\mathcal{F}$, then A is also a successful PAC learner for $\mathcal{F}$

4 Least Square Regression

4.1 Question 8

Here you will code a Matlab (or Octave) function that implements the Least Square Regression algorithm. The input to the algorithm is the training set $S = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_m, y_m)\}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, for $i = 1, \cdots, m$. The output is the hyperplane solution of the Least Square problem, $\boldsymbol{w}^* \in \mathbb{R}^d$. That is
$$\min_{\boldsymbol{w}, w_0} \sum_{i=1}^m (y_i - (w_0 + \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle))^2$$

- The prototype of the function must be

  ```
  [w,w_0] = train_ls(X,y,bias)
  ```

  where $X \in \mathbb{R}^{m \times d}$ is the matrix of m input vectors in d dimension, i.e. each input $\boldsymbol{x}_i$ is a row of X, y is a column vector of m columns containing the labels associated to the training samples. The flag 'bias' signals if a bias w_0 must be used or not. If no bias is used, the returned value of w_0 is 0. The code must be robust to the case the the matrix $X^\top X$ is not invertible.

- Implement another version, with prototype

  ```
  w=incremental_train_ls(X,y)
  ```

that construct the solution incrementally. This version does not have the bias w_0 term. This implementation must use the Sherman-Morrison formula to update the inverse of the matrix $X^\top X$:

$$(A + \boldsymbol{u}\boldsymbol{v}^\top)^{-1} = A^{-1} - \frac{A^{-1}\boldsymbol{u}\boldsymbol{v}^\top A^{-1}}{1 + \boldsymbol{v}^\top A^{-1}\boldsymbol{u}}$$

- Verify numerically that the solutions of the two algorithms on a random training set are the same.
- Discuss the computational complexity of both versions and compare them.

4.2 Question 9

In this question, you are supposed to implement linear regression using polynomial basis functions. Use only monomials of a single variable, e.g. $(x_1, x_1^2, x_1^3, \cdots, x_2, x_2^2, \cdots)$, and no cross-terms, e.g. $(x_1x_2, x_1^2x_2, \cdots)$.

- Implement a Matlab function that normalizes all the input data attributes to be between -1 and 1.

  ```
  [X_train_norm, X_test_norm] = normalizeAll(X_train, X_test)
  ```

  Note that only training data can be used for learning the normalization transformation. This will be used to prevent numerical instability in dealing with numbers that are too big or too small.

- Implement a function that generates a matrix of input samples that contains the powers of each feature from 1 to k

  ```
  [X_poly] = generate_poly_features(X,k)
  ```

- Use the training and test data provided in the file “cadata.mat”. They correspond to a random split train/test of the Housing dataset from the UCI repository. The task is to predict the median house value from features describing a town. Use the code you wrote above to generate polynomial features from $k = 1$ to $k = 5$, normalize the features, and train 5 different LS regressors (with bias w_0) with these features. Plot training error and test error for each value of k and discuss the results.