Homework 1

25/10/2018

Submission date: 20/11/18

- Submission is individual or in pairs. Make sure to write the names of both students in case the submission is in pairs.
- Submit a ZIP file containing all your files(including txt/csv files needed to run the python code) named with 9 digit of your ID and your name. Submission Example: ”200567989.zip”
- Submit your HW solution as an ‘Jupyter/IPython notebook’, after you zip your files extract them into a new folder and make sure there are no runtime errors.
- Use python version 2.7

1 Gaussian RV- Basics

1. Let $Z \sim \mathcal{N}(0,1)$ be a normal Gaussian RV, and $X \sim \mathcal{N}(\mu,\sigma)$. The Cumulative Distribution Function (CDF) of Z is defined as

   $$P(Z \leq c) \triangleq \phi(c).$$

   Express $P(X \leq x)$ using $\phi(x)$.

2. Consider a sequence of N iid RV's $\{X_i\}_{i=1}^N$, where $X_i \sim \mathcal{N}(10,1)$. The empirical mean is given by $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$. What is the distribution of $\bar{X}_N$?

3. What is the probability $P(9.5 \leq \bar{X}_N \leq 10.5)$ for $N = 1, 10, 20$? Express first using the function $\phi(x)$ and use the Python class scipy.stats.norm to calculate $\phi(x)$ and obtain a numerical value.

4. As we discussed in tutorial, since Gaussian RV is not bounded, we cannot use Hoeffding's inequality to bound terms of the form $P(9.5 \leq \bar{X}_N \leq 10.5)$. A possible alternative for this is to use the following proposition[1].

   Proposition 1. *Let $\{X_i\}_{i=1}^N$ be iid RV with $X_i \sim \mathcal{N}(\mu, \sigma)$. Then,*

   $$P(|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu| \geq \epsilon) \leq 2\exp(-\frac{N\epsilon^2}{2\sigma^2}).$$

   Use this proposition to lower bound bound $P(9.5 \leq \bar{X}_N \leq 10.5)$ for $N = 1, 10, 20$ as in previous section.

2 Parametric and Non-Parametric estimation

1. Suppose $\hat{\theta}$ is an estimator for an unknown parameter θ. Show that

   $$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}^2(\hat{\theta})$$

2. Let $X_1, .., X_N \sim$ Bernouli(p) and let $Y_1, .., Y_N \simeq$ Bernouli(q) be iid RVs. Find a plug-in estimator and estimated standard error for p. Find an approximated 90/95/99 percent confidence intervals for p. Find the plug-in estimator and estimated standard error for $p-q$. Find an approximated 90 percent confidence interval for $p - q$.

3. Let $X_1, .., X_N \sim \mathrm{B}(10, \theta)$ (Binomial distribution) iid RVs. Estimate θ using the MLE method.

4. Let $X_1, .., X_N \sim F$ be iid RVs, where F is an arbitrary, unknown CDF. Let $\hat{F}$ be the empirical distribution function. For a fixed F , use the central limit theorem to find the limiting distribution of $\hat{F}_n(x)$.

5. In the lecture and tutorial, we stated the DWK theorem and derived a C.I for the empirical CDF, for 1D type of data. Derive a C.I for the empirical CDF in a general dimension, i.e, as a function of $C(k)$(See non-parametric chapter in the lectures). If $C(k) \sim \exp(k)$ what does it mean about the hardness of the problem in high dimension?.

6. Calculate $E[\hat{F}_n]$ and $\mathrm{Var}[\hat{F}_n]$ using the definition of the empirical distribution function(remember, $X_1, .., X_N$ are iid samples from F).

[1] The proposition is more general and holds for a sum of sub-Gaussian RVs.

3 Exam question

In order to check electrical devices, a system performs repeated tests on a device until its first failure appears. The tests were performed on N devices. Denote K_i as the number of tests that were performed on the ith device(including the final test where the failure appeared), where $i \in \{1, .., N\}$. Assume that $K_1, \ldots, K_N$ are i.i.d random variables.

1. Find a non-parametric estimator(i.e, the plug-in estimator) for $E[K]$.
2. Find a non-parametric estimator for $\hat{p}_3 \triangleq P(K = 3)$, the probability a failure will occur in the third test.
3. Suggest a CI for $\hat{p}_3$ for $\alpha = 0.05$, $N = 100$.

Assume $K \sim \text{Geom}(p)$(Geometric Distribution), where p is unknown.

4. Calculate the MLE for p and for the mean $\mu = E[K]$
5. Calculate the probability that the number of tests is odd, i.e, that K is odd,P (K is odd). Simplify as far as possible.

4 Python

1. Generate 100 samples from a $N(0, 1)$ distribution. Compute a 95% CI for the CDF. Repeat this 1000 times and compute the percentage of time that the interval contained the CDF. In addition plot in a single figure the true CDF the best and the worst experiment (use $max_x|F(x) - \hat{F}_n(x)|$ as quality measure).
2. In the following we will work with the Samsung data(samsungData.csv from the course website). Read the PythonPart1.ipynb file for basics. We advise to work with the panda package(as we show in PythonPart2.ipynb). Most of the numerical computation can be performed using functions from the pandas package.
3. Compute the empirical correlation between all pairs of features. Show results in a table/heat maps.
4. Which two features are most correlated? Try to explain the results?
5. Compute the empirical correlation between all pairs of features per class. (Show results in tables/heat maps).

6. Which two features are most correlated (over all classes)? Try to explain the results?

7. Plot the convergence graphs (as was done in tutorial 2; the Bootstrap estimator value as number of samples from the data increases) of the Bootstrap estimator for the correlation variance, for the following pairs:

   - tGravityAcc-mean()-Z and tGravityAcc-min()-Z (indices 42,54).
   - fBodyAcc-bandsEnergy()-41,48.2 and fBodyAccJerk-iqr()-Z (indices 335, 365)
   - tBodyAccJerk-max()-X and fBodyGyro-skewness()-X (indices 89,454).

   Repeat for the results in 3,5 (just for the pairs listed here).

   These graphs suppose to demonstrate convergence of the Bootstrap estimator. Thus, justifying the choice of the number of time one samples from the data.

8. Using the variance estimator (for the correlation) obtained in the previous section obtain C.I with 95% on your estimators (Hint: Use Chebyshev's inequality).