Data Analysis (046193) HW2

Submission date: 29/04/17

* Submission is **in pairs only**!

* Submit a ZIP file containing your files named with 9 digit of your IDs. Submission Example: "987654321_123456789.zip".

* Submit your computer exercise solution **as an "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.**

Part 1 – Dry

Parametric estimation

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

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \left(bias(\hat{\theta})\right)^2$$

2. Let $X_1, \dots, X_n \sim Bernoulli(p)$ and let $Y_1, \dots, Y_n \sim Bernoulli(q)$. Find the 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 $X_1, \dots, X_n \sim B(10, \theta)$ (binomial distribution). Estimate θ using both MLE method and the method of moments.
4. Let $X_1, \dots, X_n \sim F$ and let $\hat{F}$ be the empirical distribution function. For a fixed x , use the central limit theorem to find the limiting distribution of $\hat{F}_n(x)$.
5. In lecture 2, C.I based for the empirical CDF using DKW theorem was introduced. For the case $k = 1$ we have that $\epsilon_n = \sqrt{\ln\left(\frac{2}{\alpha}\right)/2n}$. Derive ϵ_n for the general case as a function of $C(k)$.
6. In this problem will show that $MLE = \max_{\theta} L_n(\theta) = \min_{\theta} KL\left[\hat{f}_n(x) || f(x|\theta)\right]$. where,

$\widehat{f_n}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - X_i)$ – Empirical pdf.

$KL[f||g] = \int f(x)\log(\frac{f(x)}{g(x)})\mathrm{dx}$ - Kullback Leibler divergence.

$L_n(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$. – Likely - Likelihood function.

a. Show that $\int \widehat{f_n}(x)dx = \widehat{F_n}(x)$.
b. Show that $KL\left[\widehat{f_n}(x)\middle|\middle|f(x|\theta)\right] = g(x) - E_{\widehat{f_n}(x)}[\log f(x|\theta)]$, Where $g(x)$ is a function of x alone.
c. Show that $E_{\widehat{f_n}(x)}[\log f(x|\theta)] = \frac{1}{n}\log L_n(\theta)$.
d. Show that $\max_{\theta} L_n(\theta) = \min_{\theta} KL\left[\widehat{f_n}(x)||f(x|\theta)\right]$.

שאלה 7- מועד א' אביב 2016:

מערכת בדיקת איכות של רכיבים חשמליים מסוימים כוללת ביצוע בדיקות חוזרות על הרכיב עד להופעה ראשונה של כשל. הבדיקות בוצעו על N רכיבים. עבור $i = 1,\ldots,N$ נסמן ב- K_i את מספר הבדיקות שבוצעו על רכיב i (כולל הבדיקה שבה הופיע הכשל). אנו מניחים כי $K_1,\ldots,K_N$ הם משתנים מקריים בלתי תלויים ומפולגים באופן זהה (iid).

א. מיצאו משערך לא-פרמטרי (משערך הצבה) עבור $E(K_1)$, תוחלת מספר הבדיקות של רכיב נתון כלשהו.

ב. מיצאו משערך לא-פרמטרי (משערך הצבה) עבור $p_3 \square P(K_1 = 3)$, ההסתברות שהכשל יתרחש בבדיקה השלישית.
הנחיה: יש להראות כיצד המשערך מתקבל מתוך ההגדרה.

ג. הציעו מרווח סמך עבור המשערך $\hat{p}_3$ שמצאתם בסעיף הקודם, עבור רמת בטחון של $\alpha = 0.05$, $N = 100$.

נניח מעתה, בנוסף, כי $K_1 \square Geom(p)$ הוא בעל פילוג גיאומטרי עם פרמטר לא ידוע p .

ד. חשבו את משערך הסבירות המירבית (MLE) של הפרמטר p , ושל התוחלת $\mu_K = E(K_1)$.

ה. חשבו את משערך הסבירות המירבית של ההסתברות שמספר הבדיקות יהיה אי-זוגי: $p_{odd} = P(K_1 \text{ is an odd number})$. יש לפשט ככל האפשר.

Part 2 – Computer Exercise

1. Generate 100 samples from a $N(0,1)$ distribution. Compute a 95% confidence band for the CDF. Repeat this 1000 times and compute the percentage of time that the interval contained the CDF. In addition plot in one figure the true CDF the best and the worst experiment (use $\max_x\{|\widehat{F}_n(x) - F|\}$ as quality measure).
2. Load Samsung data from Tutorial 1.
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 <u>per class</u>. (Show results in tables/heat maps).
6. Which two features are most correlated (over all classes)? Try to explain the results?
7. Use Bootstrap method in order the estimate the variance to the empirical correlation estimators in 3, 5.
8. Plot several figures showing converges of the Bootstrap estimator to justify your choice of iteration number for the Bootstrap method.
9. Using the variance estimator obtained in the previous section obtain C.I with 95% on your estimators. (Show the new table with the C.I on the values).
10. Assume the measurement from each feature are Gaussian. Compute the MLE estimators for the Gaussian distribution per feature in the "walk" class. Which of the features best fit for the Gaussian distribution? Justify your answer.