Домашнее задание

1. Прочитайте данные из файлов train.npz и test.npz Каждый файл содержит 2 массива - samples(данные) и answers(класс).
   $samples_i \in R, answers_i \in \{0, 1, 2\}$
2. Выделите данные, соответствующие каждому классу 0,1,2 на обучающей выборке

In [ ]:

3. Визуализируйте выборку для каждого класса и сделайте предположение о виде функции распредления

In [ ]:

4. Сделайте состоятельные точечные оценки параметров распределений. (Видов распределений вам может встретиться всего 3 - равномерное, нормальное и экспоненциальное. Для равномерного $U(a, b)$ распределения сделать оценки a и b, для нормального $N(\mu, \sigma^2)$ оценки μ и σ^2, для экспоненциального $Exp(\lambda)$ оценку λ)

In [ ]:
```
def params_class0(sample):
    return
def params_class1(sample):
    return
def params_class2(sample):
    return
```

In [ ]:
```
p0 = params_class0(samples1)
p1 = params_class1(samples2)
p2 = params_class2(samples3)
```

5. Постройте графики распределений для исходной выборки и для выборки сгенерированной с параметрами, найденными с помощью точечных оценок

In [ ]:

6. Посчитайте статистику критерия Колмогорова для проверки гипотезы о том, что исходные данные являются реализацией случайной величины с функцией распределения, полученной по точечным оценкам

Пусть $D_n = \sup\limits_{-\infty<x<\infty} |\hat{F}_n(x) - F(x)|$,

где $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x\}}$ - это эмпирическая функция распределения, $F(x)$ - функция распределения полученная с помощью оценок.

Посчитайте $\sqrt{N} * D_N$, где N - число элементов выборки для каждой из выборок 1,2,3

При каком уровне значимости мы можем принять гипотезу о принадлежности выборки соответствующему распределению?

In [ ]:

7. На основе посчитанных параметров распределения постройте байесовский классификатор

In [ ]:
```
def classification_bayes(sample):
    return
```

8. Для каждого класса на тестовой выборке посчитайте количество True Positive, False Positive и False Negative.

Перед нами стоит задача бинарной классификации:

$X \to Y, Y = \{+1, -1\}$

Предположим, что мы используем алгоритм классификации $a(x_i) = y_i$

Класс с меткой "+1" называется **positive**

Класс с меткой "-1" называется **negative**

Матрица ошибок

Для классификации ответов нашего бинарного классификатора используется **матрица ошибок (confusion matrix)**:

		Правильный ответ	
		$y = +1$	$y = -1$
Выход алгоритма	$a(x) = +1$	True Positive	False Positive (Ошибка 1 рода)
	$a(x) = -1$	False Negative (Ошибка 2 рода)	True Negative

Рассмотрим теперь классификацию на более, чем 2 класса.

Для каждого класса C_j мы можем рассмотреть задачу бинарной классификации B_j:

$X \to Y, Y = \{+1, -1\}$, где $Y = +1$ если $a(x_i) = C_j$, иначе $Y = -1$

Значения True Positive, False Positive, False Negative для класса C_j есть соответствующие значения для задачи бинарной классификации B_j

In [ ]:

```
def errors(sample, correct_answer,cl):
    return
for i in range(3):
    print(i,'-',errors(samplesTest,answersTest,i))
```