Hypothesis Testing

Lesson Objectives

- After completing this lesson, you should be able to:
  - Perform hypothesis testing for goodness of fit and independence
  - Perform hypothesis testing for equality of probability distributions
  - Perform kernel density estimation

Hypothesis Testing

- Used to determine whether a result is statistically significant, that is, whether it occurred by chance or not
- Supported tests:
  - Pearson's Chi-Squared test for goodness of fit
  - Pearson's Chi-Squared test for independence
  - Kolmogorov-Smirnov test for equality of distribution
- Inputs of type `RDD[LabeledPoint]` are also supported, enabling feature selection

Pearson's Chi-Squared Test for Goodness of Fit

- Determines whether an observed frequency distribution differs from a given distribution or not
- Requires an input of type `Vector` containing the frequencies of the events
- It runs against a uniform distribution, if a second vector to test against is not supplied
- Available as `chiSqTest()` function in `Statistics`

Testing for Goodness of Fit

```
vec = Vectors.dense([0.3, 0.2, 0.15, 0.1, 0.1, 0.1, 0.05])
```

```
goodnessOfFitTestResult = Statistics.chiSqTest(vec)
```

```
goodnessOfFitTestResult.statistic
```

```
0.295
```

```
goodnessOfFitTestResult.pValue
```

```
0.999520973435643
```

```
goodnessOfFitTestResult.nullHypothesis
```

```
u'observed follows the same distribution as expected.'
```

Pearson's Chi-Squared Test for Independence

- Determines whether unpaired observations on two variables are independent of each other
- Requires an input of type `Matrix`, representing a contingency table, or an `RDD[LabeledPoint]`
- Available as `chiSqTest()` function in `Statistics`
- May be used for feature selection

Testing for Independence

```
from pyspark.mllib.linalg import Matrices
mat = Matrices.dense(3, 2, [13.0, 47.0, 40.0, 80.0, 11.0, 9.0])
```

```
independenceTestResult = Statistics.chiSqTest(mat)
```

```
independenceTestResult.statistic
```

```
90.22588968846716
```

```
independenceTestResult.pValue
```

```
0.0
```

```
independenceTestResult.nullHypothesis
```

```
u'the occurrence of the outcomes is statistically independent.'
```

Another Simple Test for Independence

```
from pyspark.mllib.regression import LabeledPoint
obs = sc.parallelize([LabeledPoint(0,Vectors.dense(1.0,2.0)),
                      LabeledPoint(0,Vectors.dense(0.5,1.5)),
                      LabeledPoint(1,Vectors.dense(1.0,8.0))])
```

```
featTestResults = Statistics.chiSqTest(obs)
```

```
map(lambda r: {r.statistic, r.pValue, r.nullHypothesis}, featTestResults)
```

```
[{0.3864762307712326,
  0.75,
  u'the occurrence of the outcomes is statistically independent.'},
 {0.22313016014843035,
  3.0000000000000004,
  u'the occurrence of the outcomes is statistically independent.'}]
```

Kolmogorov-Smirnov Test

- Determines whether or not two probability distributions are equal
- One sample, two sided test
- Supported distributions to test against:
  - normal distribution (distName='norm')
  - customized cumulative density function (CDF)
- Available as `kolmogorovSmirnovTest()` function in `Statistics`

Test for Equality of Distribution

```
data = RandomRDDs.normalRDD(sc, size=100, numPartitions=1, seed=13)
```

```
ks_result = Statistics.kolmogorovSmirnovTest(data, "norm", 0, 1)
```

```
ks_result.statistic
```

```
0.12019890461912125
```

```
ks_result.pValue
```

```
0.10230385223938121
```

```
ks_result.nullHypothesis
```

```
u'Sample follows theoretical distribution'
```

Kernel Density Estimation

- Computes an estimate of the probability density function of a random variable, evaluated at a given set of points
- Does not require assumptions about the particular distribution that the observed samples are drawn from
- Requires an RDD of samples
- Available as `estimate()` function in `KernelDensity`
- In Spark, only Gaussian kernel is supported

Kernel Density Estimation

```
from pyspark.mllib.stat import KernelDensity
```

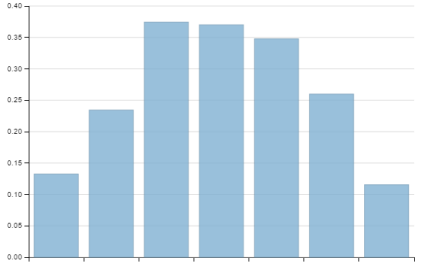
```
kd = KernelDensity()
```

```
kd.setSample(data)
```

```
kd.setBandwidth(0.1)
```

```
kd.estimate([-1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5])
```

```
array([ 0.1023487 ,  0.15699217,  0.2957955 ,  0.51760411,  0.38091952,
        0.30242779,  0.1841904 ])
```



Lesson Summary

- Having completed this lesson, you should be able to:
  - Perform hypothesis testing for goodness of fit and independence
  - Perform hypothesis testing for equality of probability distributions
  - Perform kernel density estimation