Contents

1 Pitstop extended datascience challenge

Please answer in complete sentences and show enough calculation to indicate your quantitative reasoning. Keep the total answer to 5 pages or less. *Note that there is a lot of explanation of each problem, but scripts are provided to generate the data sets.* Thanks for your interest in Pitstop, and good luck.

2 Question 1: Independence

You hear a proverb - 'red sky in the morning, sailor take warning' which suggests a 'red sky in the morning' predicts a storm. You do a simple observational study for many days and check whether there is a 'red sky' and whether there is a 'storm'. Suppose you observe:

Table 1. Red sky vs storms

	Sky	
	Red	Not Red
Storm		
yes	10	**X**
no	40	60

What value of **X** would show that 'red sky' and 'storm' were independent events? What test would you use on these data to determine if the proverb contains some truth?

3 Question 2: Testing group differences

You have an apple orchard with 50 trees and you want to test a new growth enhancer to see if it increases yield. You choose 25 trees randomly and apply the growth enhancer to them, and leave the other half alone. At the end of SEASON 1 you tabulate the total yield for each tree in pounds. Refer to the file **AppleData.dat**.

Did the treatment work? Explain. Hint: Apply an appropriate statistical test.

The next year (2) you try a different product and follow the same procedure (data for SEASON 2 is in the same file). Did the treatment work? Explain. How confident can you be about the merits of the 2 products?

4 Question 3: Flow through a hose

See Figure 1. The amount of fluid passing through a hose is a function of the hose cross-section and the pressure generated by a pump according to an input variable. A hose may accumulate internal buildup on the interior walls that narrows the cross-section of the hose. We can measure the amount of fluid passing through via a flow sensor, and want to detect when the hose is becoming blocked. The pump generates a varying pressure at one end of the hose, and the flow is proportional to the pressure drop from pump to sensor.

We do not observe the pump pressure because there is no sensor for it, but it is some curvilinear function of the RPM such as a logistic function containing an unknown constant. Assume we do observe the RPM along with the flow.

The build up of deposit mass is assumed proportional to flow rate. At the beginning when the inside of the hose is clean and a thin layer of deposits does not much affect flow. Later on, the available interior surface area is less and the hose clogs up rapidly. It will be obvious if the flow falls to 50 percent, for example; but by then the risk of a breakdown might be high, so we want an earlier warning.

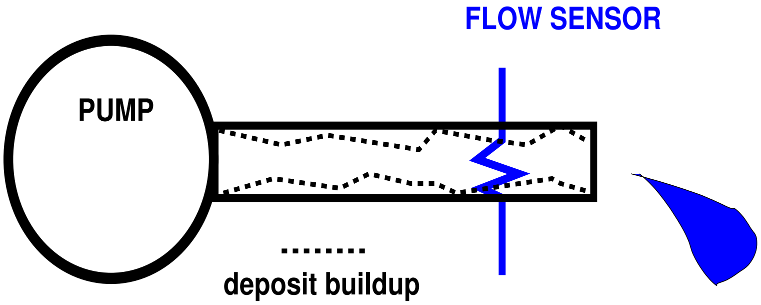

Figure 1: We want to make an early warning system to determine if a hose is clogged based on flow through the hose.

There are 2 observed variables 1) RPM (R) 2) flow (F) and two unobserved state variables of interest 1) Pump pressure (p) and 2) hose interior cross-section (S) which have to be estimated. We want to detect when fluid cross-section is reduced by 20 percent.

Let F be the fluid flow as measured by the sensor, and let R be RPM which follows a quasi-random process, let p be pump pressure and S be the hose cross-section. Denote by S_0 the threshold cross-section where a problem should be reported. We also expect to see some noise ϵ in the sensor where the ϵ_i are independently and identically distributed with zero mean expectation. Then (assuming that the pressure and RPM are related by a logistic function):

$$F = \beta p S + \epsilon \tag{1}$$

$$p = \alpha R/(1000 + R) \tag{2}$$

$$S = \gamma / \int_0^T F dt \tag{3}$$

where α,β,and $\gamma << 1$ are unknown constants. The objective is to determine S from the time series $F_i, i = 1, 2, 3, .., R_i, i = 1, 2, 3, ...$ and $\epsilon_i, i = 1, 2, 3...$ as quickly as possible. Remember that both ϵ and R are random variables. The integral equation for S may be useful for estimating part lifetime, but will not be needed for the immediate test, since S can be estimated from R and F.

It should not be difficult to estimate $K = \alpha\beta$ by substituting the second equation into the first and doing a regression of F on $R/(1 + R)$. If the critical cross-section S_0 is known, we can then divide (F, R) space into 3 regions. See Figure 2. After averaging out errors, any observations in the 'Bad' region, indicate that the hose cross-section is too small, while observations in the 'Acceptable' region indicate that the hose cross-section is big enough.

4.1 Question 3 (continued) : Machine learning approach to hose blockage

An alternate way to proceed is to simulate the data using pseudo-random inputs and make a training set for input to a neural network. While it would still be necessary to start with theoretical equations a detailed visualization would not be needed because we could generate

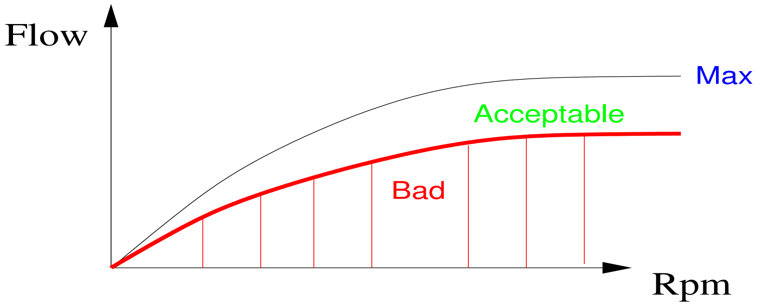

Figure 2: There is a curve in (F, R) space which separates 2 regions. It is defined by $F(1 + 1/R) < \alpha\beta S_0$. In one region, the hose has a cross-section below the critical value S_0 and in the other (acceptable) region the hose has a cross-section above the critical value. Other regions above the 'Max' curve should never be visited, though some data points may be seen there due to sensor error. In 2 dimensions it is simple to draw this curve.

simulated data points where $S < S_0$, and where $S >= S_0$ without ever visualizing the 'bad' region.

Use the script **hosesim_data.py** to generate 500 synthetic data samples of about 10 minutes with a synthetic data series sampled every 10 seconds using the model equations. The script uses matplotlib to show good and bad regions.

The data is randomly choosen episodes of 1- 10 minutes to make up the 10 minutes at a variety of fixed RPM levels, making up a total of 31290 sample points. Each 10 minute sample is generated using a different value of S, some $S < S_0$ (bad) and some $S > S_0$ (good). Every point in every sample is labelled.

After generating the data using hosesim_data.py, use SVM or other standard ML method to train model to discriminate betwee good and bad samples.

Compute the confusion matrix and error rate for the model you trained on at least 5 fresh samples. Discuss advantages and drawbacks of this approach compared to physical analysis of the hose.

5 Question 4: A simple car suspension model

Refer to Figure 3. In a second example we will consider a situation where there will be a model relationship between the variables at time T and time T-1. This is a simplified model of a car suspension which consists of a damped mass-spring system on a cart, driven over an uneven road. The objective will be to recognize those data samples where the damping parameter has fallen below a critical value.

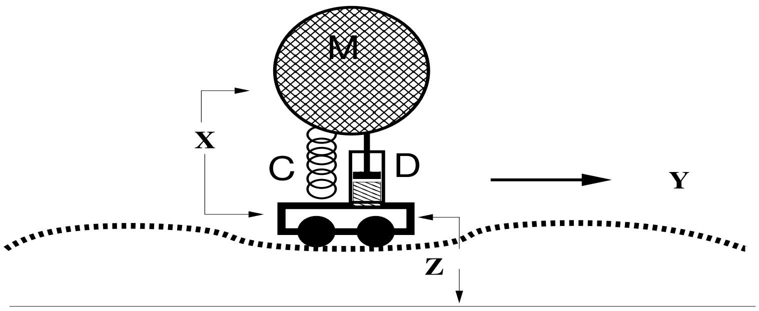

Figure 3: Cart carries a mass-spring-damper system over an uneven road with known profile $Z(Y)$ with a known time profile $Y(t)$. The mass will undergo motions in response to the unevenness in the road.

5.1 Damping

If X is the position of a mass on a spring + damper with the spring at equilibrium for $X = 0$, the equation for X is:

$$f(t) = M\ddot{X} + D\dot{X} + CX \tag{4}$$

where $f(t)$ is a force input, t is time, M is mass, $D > 0$ is a viscosity parameter and C is a spring constant. Direct substitution of $X = exp(\lambda t)$ where $F = 0$ and solving the resulting quadratic equation gives: $\lambda = (-D \pm \sqrt{D^2 - 4MC})\ /\ 2M$, which implies that any sinusoidal oscillations die out at a rate proportional to $exp(Re(\lambda))t)$. The coefficient $D/2M$ can be called the damping coefficient. When it is too small, oscillations may persist and cause problems such as loss of traction. This is a particular problem if the periodic force input frequency matches a resonant frequency of the system.

5.2 Cart equations

Assume we can observe Z the position of the cart above sea level, and X, the vertical position of the mass relative to the cart and Y the position of the cart along the horizontal track. Assume that the road has a known profile $Z(Y)$ and that the cart is driven with a known position function $Y(t)$. For simplicity let $Y = Vt$ where V is a known speed. Let X_0 be the equilibrium position of the spring and assume that restoring forces are symmetric with respect to displacement above and below the equilibrium point. Then the position of the mass above sea level is $P = X + Z$, and the mass will bounce up and down in response to vertical forces as the cart moves. Then the equations:

$$M\ddot{P} = -C(X - X_0) - D\dot{X} \tag{5}$$

$$P = X + Z(Y(t)) \tag{6}$$

$$Y(t) = Vt \tag{7}$$

describe the motions of the cart and the motion of the mass relative to the cart. The cart wheels need to stay in contact with the ground for these equations to be true ie. the upward

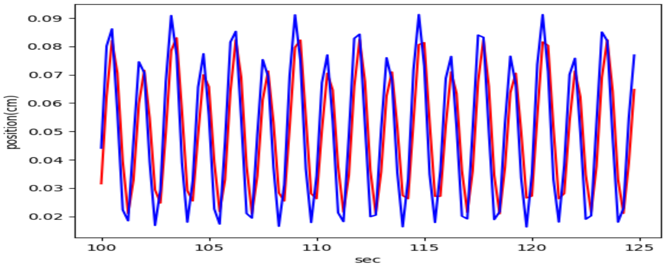

Figure 4: Cart with D=5000, (red) versus cart with D=500 (blue) response to the same road according to the model difference equations. The mass position time-series was not very sensitive to the value of D in the sense that the two time-series showed small steady state differences. However the picture shows the response after the initial transient. In a linear system, outputs will eventually be at the input frequency for periodic inputs.

force exerted by the rising mass should not exceed the weight of the cart. For now we assume it is heavy enough that this never happens.

5.3 Difference equation model

The physical equations were modelled using difference equations as follows:

$$LHS = M(X_{n+1} - 2X_n + X_{n-1})/\delta T^2 + D(X_{n+1} - X_{n-1})/2\delta T + CX_n \tag{8}$$

an approximation which yielded stable solutions, where $LHS = -M\ddot{Z} + CX_0$, where $\ddot{Z}$ was obtained by directly differentiating a model of the road roughness based on fourier components. This model was:

$$Z = \sum_{K=1/2,1,2,4} a_K\ sin(KPY) + b_K\ cos(KPY) \tag{9}$$

where $Y = Vt$ and assuming the vehicle moves at $60km/hour$ taking $P = 16.66m$ implies a lowest roughness driving frequency at 0.5 Hz. We took a_k and b_k to be random coefficients chosen from $0.01, 0.02, 0.025, 0.03, 0.035$ ie. amplitudes ranging from 1cm - 3.5 cm. $M = 2000kg$, $C = 0.6 \times 10^4$, $X_0 = 0.05m$, $\delta T = 0.25$ sec. The values of D were taken as $(500, 600, 700, 800, 400, 300, 200)$ (low) or $(5000, 5500, 6000, 6500, 4500, 4000, 3500)$ (high).

5.4 Question 4: (continued) Machine learning discrimination of suspension quality

Fit machine learning models to multiple samples of 25 seconds of data with either high or low values of D. Use the script **cartsim_data.py**, which is provided. The script contains directions for producing data for the different cases.

You will do 100-1000 runs of 500 time points (125 seconds), using the last 25 seconds of each run. The objective is to discriminate between data taken from runs with low values of D (bad) and the data taken from runs with high values of D (good).

There were 2 types of observation series. In the first type of observation (**road input**) we fitted tuple data of the form $[LHS, X_{n-2}, X_{n-1}, X_n]$. This assumes we can observe the actual vertical forces generated by the roughness in the road along with the mass position response at 3 adjacent timepoints 0.25 seconds apart.

In the second type of simulation study (**in vehicle vibration**) we tried to discriminate D values using $[X_{n-2}, X_{n-1}, X_n]$ or longer sub-series such as $[X_{n-5}, X_{n-4}, ...X_{n-1}, X_n]$, (model VV5) and $[X_{n-8}, X_{n-7}, ...X_{n-1}, X_n]$ (model VV8), so that direct knowledge of the road profile was not assumed. This situation will be typical of the real world, since no independent measurements of road profiles will be commonly available for the length scales of interest.

There were 2 types of experiment. The first type of experiment involved datasets where vehicles were assumed to drive over a single standard rough track (**one standard road**), and the second type of experiment used datasets where vehicles were assumed to drive over a fresh and different rough track (**random roads**) on each simulated realization. While these experiments assume the vehicle is driven at a certain speed, it is likely that independent speed measurements will be available in the field, so that real data sets could be normalized for speed.

Models were fitted on 70 percent of the randomly shuffled dataset (N runs x 100 points) while the remaining 30 percent was used to evaluate success using the confusion matrix and overall error rate (number good when bad + number bad when good / total N).

6 Results Table to be filled in

We define success for a machine learning model here as an error rate consistently less than 20 percent. Generally the most difficult discrimination was the combination of **random roads** and **road input**. The easiest discrimination was **one standard road** and **in vehicle vibration**.

Fill in the following table, with 4 replications for each case. Compute the confusion matrix in each case. Discuss the advantages and disadvantages of each ML approach. Why do you think some cases are easier? Explain.

Table 2. Study results

ML algorithm	N runs	experiment type	observation type	error rates	result
Your choice	100	random roads	road input		
Your choice	200	random roads	road input		
Your choice	200	one standard road	road input		
Your choice	200	random roads	in vehicle vibration		
Your choice	200	one standard road	in vehicle vibration		
Your choice	500	random roads	road input		
Naive Bayes	200	random roads	road input		
Naive Bayes	200	one standard road	in vehicle vibration		
Naive Bayes	200	random roads	in vehicle vibration		
Naive Bayes	1000	random roads	road input		
MLP(20,5)	200	random roads	road input		
MLP(20,5)	500	random roads	road input		
MLP(20,5)	1000	random roads	road input		
MLP(20,5)	500	one standard road	road input		
MLP(20,5)	500	one standard road	in vehicle vibration		
MLP(20,5)	500	random roads	in vehicle vibration		