ECE 592 – Topics in Data Science

Test 1: Models – Fall 2020

September 9, 2020

Please remember to justify your answers carefully.

Last name: _____ First name: _____

Please recall the course academic integrity policy for tests:
No cooperation or "collaboration" between students is allowed. Especially during an online course experience, it could be tempting to text or email a friend. This is not allowed. You will be allowed to use your notes, books, a browser, and software such as Matlab and/or Python.[1] However, while working on the test you should not text, email, or communicate with other people (certainly not other students) in any way, unless you are consulting with the course staff. **By submitting the test, you will be acknowledging that you completed the work on your own without the help of others in any capacity.** Any such aid would be unauthorized and a violation of the academic integrity policy.

[1] You can use the browser to access Moodle, the course webpage, and look up technical topics. Similar to a normal test, you must not communicate with other people.

Question 1 (Probability and Bayes Theorem)
Consider two events. Under event $E1$, a student prefers to go to the beach for a vacation. Under event $E2$, a student prefers to go to the mountains for a vacation. The probabilities of these events satisfy $\Pr(E1) = 0.4$ and $\Pr(E2) = 0.7$. Moreover, the probability that a student prefers neither the beach nor the mountains is $\Pr\big((E1)^C, (E2)^C\big) = 0.2$. Please compute the following probabilities. (Note that $(\cdot)^C$ denotes the complement of a set or event.)

(a) $\Pr\big((E1)^C\big)$.

(b) $\Pr\big((E2)^C\big)$.

(c) $\Pr(E1 \cup E2)$

(d) $\Pr(E1, E2)$.

(e) $\Pr\big((E1)^C, E2\big)$.

(f) $\Pr\big(E1, (E2)^C\big)$.

Question 2 (Model Complexity)
Consider a model for text, where the alphabet is comprised of C characters, and each character is predicted by the previous 2 characters. That is, X_n has a probability mass function, $P(X_n|X_{n-1}, X_{n-2})$. We want to learn these probabilities from the data. Please describe the model complexity for length-N input strings as a function of C and N.

Question 3 (Curve fitting.)
Consider a sequence of N measurements generated as follows,

$$y = \sin(x) + \mathcal{N}(0, \sigma^2),$$

where the noise is Gaussian with zero mean and variance $\sigma^2 = 1$. The inputs x and outputs y are each of length N. Our goal in this question is to estimate the order of a reasonable Taylor approximation for the function, $\sin(x)$.
For three signals of different lengths ($N_1 = 10^2, N_2 = 10^4, N_3 = 10^6$) and several model orders, we fit coefficients to a polynomial approximation,

$$y(x, (a_i)) = \sum_{i=0}^{\text{model order}} a_i x^i.$$

Below, we list for several model orders the corresponding total squared error (TSE) for the three signal lengths (N_1, N_2, and N_3),

$$\text{TSE} = \sum_{n=1}^{N} \left(y_n - y(x_n, (a_i))\right)^2,$$

where for each model order and signal length we optimized the (a_i) coefficients.

Order	TSE (N1)	TSE (N2)	TSE (N3)
0	97.94	14205.78	1430238.64
1	81.73	10565.03	1062357.60
2	79.79	10564.51	1062351.54
3	74.77	10004.89	1000698.87
4	74.65	10004.34	1000697.44
5	73.97	9984.02	997574.81
6	73.47	9983.74	997574.29
7	73.10	9983.73	997518.10
8	71.49	9981.06	997514.20
9	71.44	9979.12	997506.23

For each signal length ($N_1 = 10^2, N_2 = 10^4, N_3 = 10^6$), which model order seems best? Are your answers identical, different? What might that imply about the amount of data (N) needed to learn?
To solve this, we suggest that you minimize the length of a two part code, where the length of describing the data given the polynomial model obeys

$$\begin{aligned}
\text{len}(\text{data}|\text{model}) &= -\sum_{n=1}^{N} \log_2(f(\text{unexplained}_n)) \\
&= -\sum_{n=1}^{N} \log_2 \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{\left(y_n - y(x_n, (a_i))\right)^2}{2\sigma^2} \right\} \right) \\
&= \frac{N}{2} \log_2(2\pi\sigma^2) + \log_2(e) \sum_{n=1}^{N} \frac{\left(y_n - y(x_n, (a_i))\right)^2}{2\sigma^2} \\
&= \frac{N}{2} \log_2(2\pi\sigma^2) + \frac{\log_2(e)}{2\sigma^2} \text{TSE},
\end{aligned}$$

and $\frac{\log_2(e)}{2\sigma^2} \approx 0.72$. (You may assume that each parameter requires $\text{par}(N) = 2 + \frac{1}{2}\log_2(N)$ bits, and so $\text{par}(N_1) \approx 5.3$, $\text{par}(N_2) \approx 8.6$, and $\text{par}(N_3) \approx 12.0$ bits.) Make sure to justify your answer.