CMPE 343: Introduction to Probability and Statistics for Computer Engineers

Fall 2024

Ali Ayhan Günder 2021400219 — Berat sayın 2019400261 — Yusuf Suat Polat 2021400312 Drive Link: Documents

Due by: December 15, 2024, 23:59

Task 1: Language Models (30 pts)

In recent years, Large Language Models (LLM), such as ChatGPT, Claude, and Llama, have emerged as powerful models that can solve many tasks using language, often reaching human performance. This achievement of LLMs can be attributed to the immense scaling of three components: computational power, model capacity, and data. Graphical Processing Units (GPU) allow us to train LLMs much faster than CPUs by enabling applying matrix operations efficiently. Also, with the increased memory of GPUs, we can increase our model's capacity to billions of parameters. Lastly, billions and trillions of gigabytes of text data are available on the internet to train these LLMs. Here is a short intro to LLMs: https://www.youtube.com/watch?v=zjkBMFhNj_g

Despite the complexity of developing LLMs with huge computational power, model size and big data, once trained, LLMs operate on a very simple mechanism to generate text: Given an input text, predict the next word over and over until we decide to stop the process or the model predicts a special word (or token) called "`<|end|>`" (similarly the model prepends another special word called "`<|start|>`" to each input text.). At each step, given the previous words, the LLM generates a probability distribution over all possible words in its vocabulary. Then, we can either obtain the word with the highest probability or we can apply random sampling from this distribution, which is the approach that is usually followed and enables LLMs to generate diverse texts.

In this task, you will develop a simple, probabilistic language model that can generate sentences by sampling the next word using conditional probabilities. For example, given the sentence "`she paints beautiful pictures`", the generation process for each word $w_i, i = 1, \ldots, 4$, is as follows:

1. $w_1 \sim P(w \mid$ `"<|start|>"`$)$, $w_1 =$ `"she"`
2. $w_2 \sim P(w \mid$ `"<|start|>"`, `"she"`$)$, $w_2 =$ `"paints"`
3. $w_3 \sim P(w \mid$ `"<|start|>"`, `"she"`, `"paints"`$)$, $w_3 =$ `"beautiful"`
4. $w_4 \sim P(w \mid$ `"<|start|>"`, `"she"`, `"paints"`, `"beautiful"`$)$, $w_4 =$ `"pictures"`

Here, $P(w \mid$ `"<|start|>"`$)$ denotes the probability of w being the first word of a sentence and $P(w \mid$ `"<|start|>"`, `"she"`$)$ denotes the probability of w coming after the first word `"she"` and so on.

Although the next word may depend on all of the previous words, it is hard to correctly estimate $P(w_4 \mid w_1, w_2, w_3)$ because it requires a lot of data (we need to find a lot of sentences that start with w_1, w_2, w_3). Therefore, we can make a simplifying assumption that each word depends only on the previous word, meaning that $w_4 \sim P(w \mid w_3), w_3 \sim P(w \mid w_2)$. You are given a `sentences.txt` as a sample of 1000 sentences in English. Using this sample, you need to calculate the conditional probability $P(w_2 \mid w_1)$ (probability that word w_2 comes after word w_1) for each pair of words w_1 and w_2.

Hint: Let us take $P($`"enjoy"` $\mid$ `"You"`$)$ for example. Think of the following: How many `"You"` are there in `"sentences.txt"`? How many `"enjoy"` are there, and how many of them come after `"You"`?

Question 1

After obtaining the probabilities, answer the following questions: Generate five sentences using the conditional probabilities that you have obtained (Each sentence needs to start with the word "`<|start|>`"). Are these sentences meaningful? Are they already in the `sentences.txt` or completely new?

Solution 1

The sentences I generated are:

- You walk in the garden
- We have a picnic in the morning walk in the café
- They play the library
- They enjoy a fitness workshop
- They practice drawing every morning coffee shop

Even though all the sentences have the same structure as the ones in the given document, only some of the sentences are meaningful. This indicates the approach that predicts the next word according to the current word only does not keep the meaning in the sentence as a whole.

None of the sentences are in the sentences.txt document.

Question 2

Suppose that we want to calculate the probability, $P(w_1, w_2, \ldots, w_k)$, of a sentence $w_1, w_2, \ldots, w_k$ with k words. Given our assumption that each word depends only on the previous word, what is the formula for $P(w_1, w_2, \ldots, w_k)$?

Solution 2

Given the assumption that each word w_i depends only on the previous word w_{i-1}, the joint probability of the sentence $P(w_1, w_2, \ldots, w_k)$ can be expressed as:

$$P(w_1, w_2, \ldots, w_k) = P(w_1) \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_2) \cdots P(w_k \mid w_{k-1})$$

In general, this can be written compactly as:

$$P(w_1, w_2, \ldots, w_k) = P(w_1) \prod_{i=2}^{k} P(w_i \mid w_{i-1})$$

Explanation

1. $P(w_1)$: The probability of the first word w_1 starting the sentence.
2. $P(w_i \mid w_{i-1})$: The conditional probability of word w_i given the previous word w_{i-1}.
3. $\prod_{i=2}^{k}$: The product notation is used to multiply the probabilities for each word transition from w_2 to w_k.

This formula computes the joint probability of the entire sequence of words in the sentence based on the assumption that each word depends only on its immediate predecessor.

Question 3

Generate a sentence, $w_1, w_2, \ldots, w_k$ using your language model and calculate $P(w_1, w_2, \ldots, w_k)$. Then, think of a sentence $v_1, v_2, \ldots, v_k$ on your own and calculate $P(v_1, v_2, \ldots, v_k)$. Which one is larger, $P(w_1, w_2, \ldots, w_k)$ or $P(v_1, v_2, \ldots, v_k)$? Comment on the results (You can calculate $\log P(w_1, w_2, \ldots, w_k)$ instead of $P(w_1, w_2, \ldots, w_k)$ to avoid numerical underflow.)

Solution 3

1. The Chosen Sentence: "I watch the clouds"

Conditional Probabilities:

- $P(I \mid \texttt{|start|}) = 0.2710$
- $P(watch \mid I) = 0.0185$
- $P(the \mid watch) = 0.4286$
- $P(clouds \mid the) = 0.0043$
- $P(\texttt{|end|} \mid clouds) = 1.0000$

Joint Probability:

$$P(\texttt{|start|}, I, watch, the, clouds, \texttt{|end|}) = 0.2710 \times 0.0185 \times 0.4286 \times 0.0043 \times 1.0000 = 0.00000912$$

Log Probability:

$$\begin{aligned}\log P(\texttt{|start|}, I, watch, the, clouds, \texttt{|end|}) &= \log(0.2710) + \log(0.0185) + \log(0.4286) \\ &\quad + \log(0.0043) + \log(1.0000) \\ &= -1.3098 - 3.9882 - 0.8453 - 5.4463 + 0 \\ &= -11.5896\end{aligned}$$

2. The Model-Generated Sentence: "You share your dog"

Conditional Probabilities:

- $P(You \mid \texttt{|start|}) = 0.2540$
- $P(share \mid You) = 0.0984$
- $P(your \mid share) = 0.6471$
- $P(dog \mid your) = 0.0714$
- $P(\texttt{|end|} \mid dog) = 1.0000$

Joint Probability:

$$P(\texttt{|start|}, You, share, your, dog, \texttt{|end|}) = 0.2540 \times 0.0984 \times 0.6471 \times 0.0714 \times 1.0000 = 0.00115546$$

Log Probability:

$$\begin{aligned}\log P(\texttt{|start|}, You, share, your, dog, \texttt{|end|}) &= \log(0.2540) + \log(0.0984) + \log(0.6471) \\ &\quad + \log(0.0714) + \log(1.0000) \\ &= -1.3698 - 2.3188 - 0.4344 - 2.6425 + 0 \\ &= -6.7655\end{aligned}$$

3. Comparison of Results

Joint Probabilities:

$$P(\text{I watch the clouds}) = 0.00000912, \quad P(\text{You share your dog}) = 0.00115546$$

Log Probabilities:

$$\log P(\text{I watch the clouds}) = -11.5896, \quad \log P(\text{You share your dog}) = -6.7655$$

The model-generated sentence, "You share your dog," has a higher probability.

4. Comments on Results

The model-generated sentence, "You share your dog," is more probable based on the given probabilities. This is expected since:

- The model likely favors more frequent word transitions observed in the training data, such as "You → share" or "share → your."
- The chosen sentence, "I watch the clouds," involves less frequent transitions, such as "watch → the" and "the → clouds," which result in lower probabilities.

The higher probability of the model-generated sentence suggests it aligns better with the training corpus.

Part 2: A Scientist on Dune

"Dune" is a science fiction franchise set on a desert planet called Arrakis, often simply called Dune. The planet is covered in endless sand dunes and is the only known source of a very valuable substance called spice that extends life and enhances mental abilities. Dune is also home to giant creatures called sandworms. These massive worms are also an essential factor in the production of spice. Sandworms live beneath the surface and are extremely sensitive to rhythmic vibrations like footsteps. If someone walks in a regular pattern, it can attract a sandworm, which will rise from the sand and destroy everything in its path.

The people of Dune, called the Fremen, have learned how to survive alongside these dangerous creatures. To avoid being detected by the sandworms, the Fremen have become accustomed to walking in an irregular, non-rhythmic way.

Assume that you are a scientist on Dune who studies these sandworms. Specifically, you are interested in the sandworms' ability to detect rhythmic vibrations. You carry out several experiments that investigate their attraction mechanisms with respect to several factors.

Detection Time

You are interested in analyzing the time it takes until a sandworm detects rhythmic movements. You have been given a device that is put on the sand and produces vibrations in a certain rhythm when activated. When the device is activated and vibrations have begun, a sandworm arrives at the device's location after a certain time t. You will analyze how this arrival time changes with respect to the complexity of the rhythmic movement. You suspect that some rhythms are harder to detect for sandworms.

Suppose that a very simple rhythm, R_1 (4/4, for example), can be detected in 25 minutes on average ($\mu_1 = 25$). You have come up with another rhythm, R_2. You have tested R_2 by independently placing the device in certain locations. You have repeated this experiment five times and obtained the following results:

Arrival Time
27.5
32.5
20.1
42.5
38.1

Table 1: Experiment results for R_2

Answer the following questions:

1. Which probability distribution would be appropriate to model the arrival time of a sandworm? What are the mean and the standard deviation of this distribution? Evaluate them for R_1.
   Answer: The arrival time can be modeled by exponential distribution, defined by a single parameter λ (rate parameter).

   - The mean of the exponential distribution is $\mu = \frac{1}{\lambda}$.
   - The standard deviation is also $\sigma = \frac{1}{\lambda}$.

   Parameters for R_1:

   $$\mu_1 = 25 \quad \Rightarrow \quad \lambda_1 = \frac{1}{\mu_1} = \frac{1}{25} = 0.04$$
   $$\sigma_1 = 25$$

   Parameters for R_2:

   $$\mu_2 = \frac{\sum x_i}{n} = \frac{27.5 + 32.5 + 20.1 + 42.5 + 38.1}{5} = 32.14$$
   $$\lambda_2 = \frac{1}{\mu_2} = \frac{1}{32.14} \approx 0.031$$

2. Conduct a hypothesis test to analyze whether the rhythmic movement R_2 results in an average detection time higher than $\mu_1 = 25$ or not. Use significance level $\alpha = 0.05$. Write out all the steps, including the hypothesis, test statistic, computations, and decision.

   Answer:

   We aim to test whether the mean detection time for R_2 is greater than 25 minutes ($\mu_1 = 25$). **Hypotheses**:

   $$H_0 : \mu_2 \leq 25 \quad (\lambda_2 \geq 0.04)$$

   $$H_1 : \mu_2 > 25 \quad (\lambda_2 < 0.04)$$

   Test Statistic: For an exponential distribution, the sample mean is used as the test statistic:

   $$Z = \sqrt{n} \cdot \frac{\bar{X} - \mu_0}{\mu_0}$$

where $\bar{X}$ is the sample mean, $n = 5$, and $\mu_0 = 25$.
Calculation: Substituting the values:

$$Z = \sqrt{5} \cdot \frac{32.14 - 25}{25} = \sqrt{5} \cdot \frac{7.14}{25} = 0.6388$$

Critical Value: At $\alpha = 0.05$, the critical value for a one-tailed test from the standard normal table is:

$$Z_{0.05} = 1.645$$

Decision: Since $Z = 0.6388 < 1.645$, we fail to reject H_0.
Conclusion: There is insufficient evidence to conclude that the mean detection time for R_2 is greater than 25 minutes. The hypothesis test does not support the claim at the $\alpha = 0.05$ significance level.

Predicting Detections

You want to carry out another study that seeks to answer which environmental settings sandworms can detect a movement. You suspect that the distance to the nearest sandworm and the amplitude (loudness) of the vibrations are crucial in the attraction of sandworms. Assume that you have done 100 experiments with varying vibration amplitudes $a_i, i = 0, 1, \ldots, 100$ and distances to the nearest sandworm $d_i, i = 0, 1, \ldots, 100$. For each experiment, you have observed whether a sandworm arrived or not in a fixed time threshold t. The results are in the `detection_data.csv`. Using this data, you will calculate the probability of detection and non-detection, given a certain amplitude a and distance d. In other words, you will compute $P(\text{Detect} \mid a, d)$ and $P(\text{No Detect} \mid a, d)$. Assume that a and d are normally distributed and **conditionally independent** given the outcome (Detect or No Detect). You can use the dataset to estimate parameters of a and d under detection and non-detection settings. You will decide that a sandworm will detect the movement at distance d and amplitude a if $P(\text{Detect} \mid a, d) > P(\text{No Detect} \mid a, d)$ and vice versa.

Answer the following questions:

1. Write out the formulas for $P(\text{Detect} \mid a, d)$ and $P(\text{No Detect} \mid a, d)$.

   Answer: The detection mechanism is modeled using the Gaussian probability density function:

   $$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

   By using Bayes' theorem and assuming a and d are conditionally independent we can reach these formulas:

   (a) **Probability of Detection:**

   $$P(\text{Detect} \mid a, d) = \frac{P(a \mid \text{Detect}) \cdot P(d \mid \text{Detect}) \cdot P(\text{Detect})}{P(a, d)}$$

   (b) **Probability of Non-Detection:**

   $$P(\text{No Detect} \mid a, d) = \frac{P(a \mid \text{No Detect}) \cdot P(d \mid \text{No Detect}) \cdot P(\text{No Detect})}{P(a, d)}$$

(c)

$$P(a, d) = P(a, d \mid \text{Detect}) \cdot P(\text{Detect}) + P(a, d \mid \text{No Detect}) \cdot P(\text{No Detect})$$

2. Make predictions for the 100 observations that you have obtained. How many of the 100 outcomes have you detected correctly?

   Answer: The model predicts detection for each observation based on:

   $$P(\text{Detect} \mid a, d) > P(\text{No Detect} \mid a, d) \implies \text{Detect}$$

   After running the model on detection data csv document, the number of correctly predicted outcomes is: 90

   Accuracy: 90%

3. Coincidentally, another scientist has conducted the same experiment as you and sent their results to you so that you can test your prediction mechanism. Keeping your estimates of distribution parameters the same, make predictions for the new set of results that are available in `detection_data_extra.csv`. How many of the outcomes did you correctly predict? Did the number of correctly predicted outcomes change when compared to your correctly predicted outcomes for your original experiment results?

   Answer: Yes, the number of correctly predicted outcomes decreased. When predictions are made on the detection data extra csv dataset, the number of correctly predicted outcomes is: 82

   Accuracy: 82%

   Compared to the original dataset, the accuracy decreased slightly. This could be due to differences in the distribution of 'Amplitude' and 'Distance' in the new dataset.

Part 3: Data Center Manager (30 pts)

Assume that you are a data center manager responsible for preventing and fixing server failures as efficiently and fast as possible. You want to recruit a set of data center operators who can fix server failures, but you are uncertain of the optimal number of employees for this task. You plan to choose the number of employees based on the average failures per day. You know that the average number of failures per day depends on the following interrelated quantities: average temperature, average server load, and average cooling efficiency. These variables influence each other to some degree. For example, server load influences the temperature since more CPU utilization would result in the production of more heat. Similarly, if the cooling efficiency is low, the average temperature in the data center would increase, and so on.

We can model these variables as a *multivariate normal distribution*. Multivariate normal distribution of three variables is written as $\mathcal{N}(\mu, \mathbf{\Sigma})$ where:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$$

denotes the mean vector and the covariance matrix, respectively.

μ_i is the mean of the variable i and σ_{ij} is the covariance between variables i and j. The probability density function of $\mathcal{N}(\mu, \mathbf{\Sigma})$ is written as:

$$f(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{3/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)$$

where $\mathbf{x}$ is a vector of random variables:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

In our case, assume that x_1, x_2, and x_3 denote the average temperature, the average server load, and average cooling efficiency, respectively. Further, you know the mean and covariance of these variables:

$$\mu = \begin{bmatrix} 20 \\ 0.3 \\ 0.8 \end{bmatrix}, \ \mathbf{\Sigma} = \begin{bmatrix} 4 & 0.5 & 0.2 \\ 0.5 & 0.7 & 0.2 \\ 0.2 & 0.2 & 0.1 \end{bmatrix}$$

Based on your past experiences, once the average temperature, the average server load, and the cooling efficiency are observed, you can determine the number of failures per day as follows:

$$g(\mathbf{x}) = 0.1x_1^2 + 12.5x_2^2 - 7.5x_3^2$$

where $\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$. You wish to calculate the expected value of $g(\mathbf{x})$ so that you can decide on the optimal number of employees.

You are tasked with the following:

1. Write the formula for $E[g(\mathbf{x})]$ but do not evaluate it.

   Answer:

   $$E[g(\mathbf{x})] = \int_{-\infty}^{\infty} g(x) f(x; \mu, \Sigma)\, dx$$

   So;

   $$E[g(\mathbf{x})] = \int_{-\infty}^{\infty} (0.1x_1^2 + 12.5x_2^2 - 7.5x_3^2)\left(\frac{1}{(2\pi)^{3/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)\right) dx$$

2. Estimate $E[g(\mathbf{x})]$ using Monte Carlo sampling with the following number of samples: $n_1 = 50$, $n_2 = 100$, $n_3 = 1000$, $n_4 = 10000$. For each experiment, calculate the 95% confidence interval.

   Answer: To estimate $E[g(\mathbf{x})]$, first we create samples using multivariate normal distribution formula. Then, we calculate the number of failures $g(x)$ for each sample and take average of them. After that, for each n we calculate the confidence interval by the formula:

   $$\overline{g} \pm z_{a/2}\sigma/\sqrt{n}$$

   Since $a = 0.05$, $z_{a/2} = 1.96$ by critical values table.

3. Estimate $E[g(\mathbf{x})]$ using Monte Carlo sampling with $n_0 = 10000$ samples and $n_1 = 50$ samples and denote your estimates as g_0 and g_1, respectively. Given that you are confident with your estimate g_0, test your estimate g_1 on whether $g_0 = g_1$ or not. Use significance level $\alpha = 0.05$. Provide all the steps, including the hypothesis, test statistic, computations, and decision.

Answer: We create our hypotheses as:

$$\text{Null Hypothesis: } H_0 : \mu_0 = \mu_1$$

$$\text{Alternative Hypothesis: } H_1 : \mu_0 \neq \mu_1$$

Then, we calculate test statistic by using two-sample z-test:

$$z = \frac{(\overline{g_0} - \overline{g_1}) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}}}$$

After that, we compare the result to the critical z-value, which is 1.96 since $a = 0.05$. If $z > 1.96$, we reject H_0. If it is other way around, we fail to reject H_0, since the difference is not significant.