باسمه تعالی

نام استاد: دکتر فاطمی‌زاده

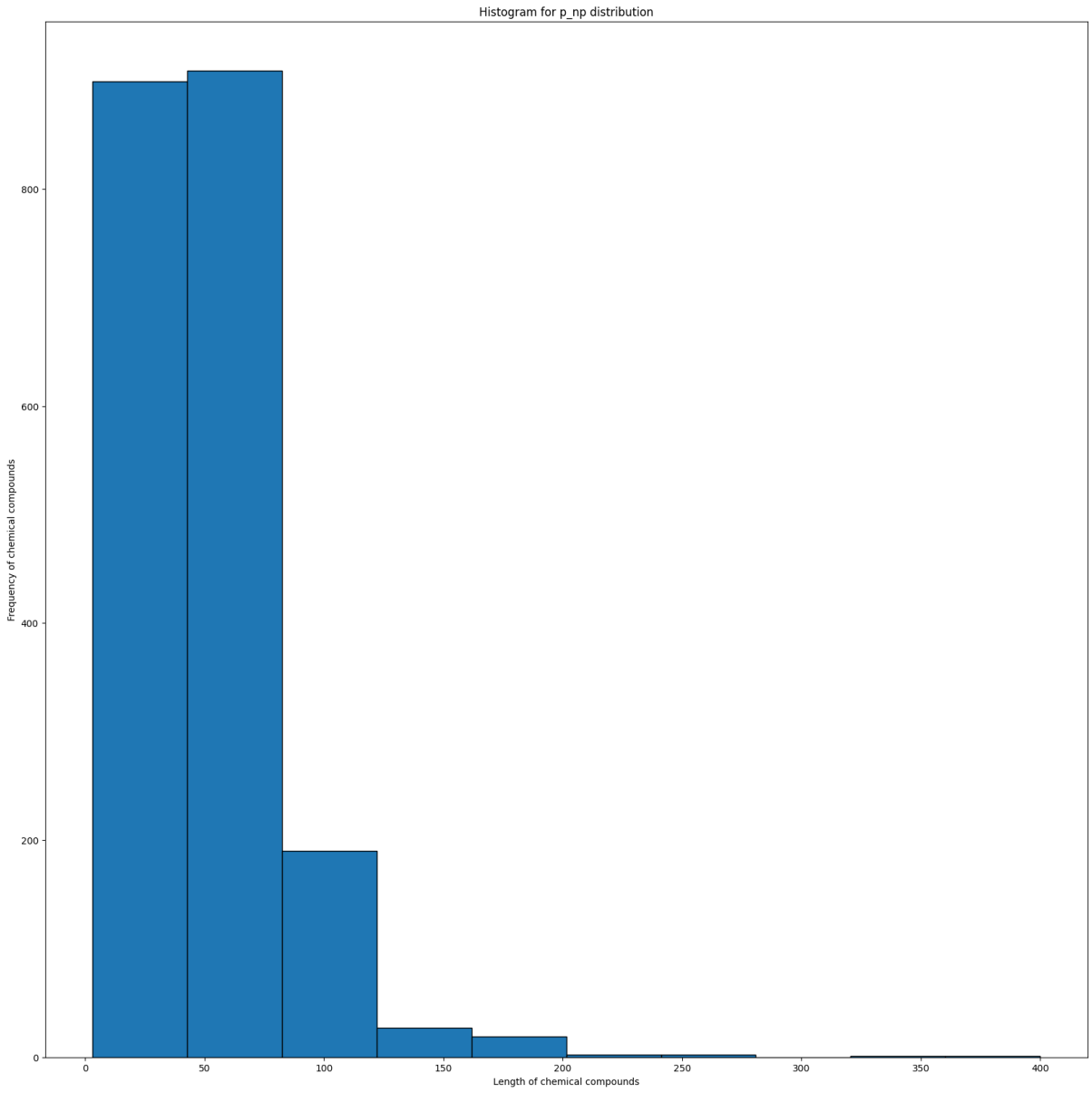
موضوع: پاسخ قسمت تئوری سوال اول از تمارین سری چهارم

تهیه‌کننده: نوید فرمهینی فراهانی

زمستان ۱۴۰۲

قسمت آ

طبق نمودار زیر، توزیع طول هر یک از smile ها، قبل از tokenize شدن، به صورت زیر است:



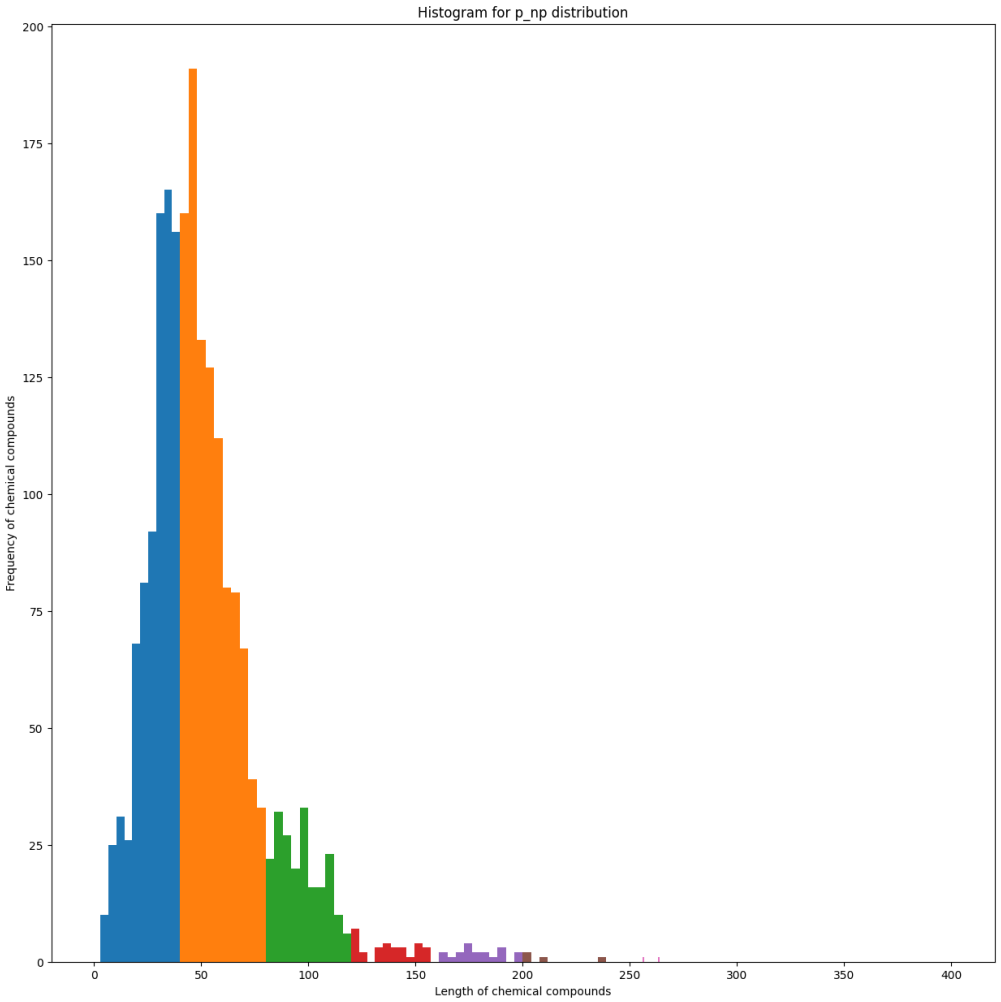
شکل ۱-۱- نمودار توزیع طول رشته‌های هر smile(تعداد آن‌ها ۲۰۵۰ است، لذا انتگرال منحنی فوق ۲۰۵۰ است)

ضمناً سایر مشخصات آماری به صورت زیر هستند:

```
For the smiles, the maximum length is 400.0
For the smiles, the minimum length is 3.0
For the smiles, the mean length is 51.47414634146342
For the smiles, the standard deviation of lengthes is 30.61318964709212
```

شکل ۱-۲- مشخصات آماری smile ها

همانطور که از شکل ۱-۱ مشخص است، اکثر این فرمولهای شیمیایی دارای طول کم هستند و برای طول بین ۲۶۰ تا ۳۲۲ هیچ گونه فرمول شیمیایی وجود ندارد. اکنون میخواهیم این منحنی ها را به صورت جداگانه در چندین بازه رسم کنیم. تعداد بازه‌های در نظر گرفته شده، ۱۰ است(هر یک از این دسته‌ها به صورت جداگانه داخل فایل note book آورده شده است).



شکل ۱-۲- نمودار توزیع طول رشته‌های هر smile؛ این توزیع به صورت یکنواخت نمی‌باشد و احتمالا توزیع token ها نیز یکنواخت نخواهد بود؛ لذا برای قسمت آخر سوال اول، نباید برای هر bin به صورت یکنواخت نمونه برداری کنیم.

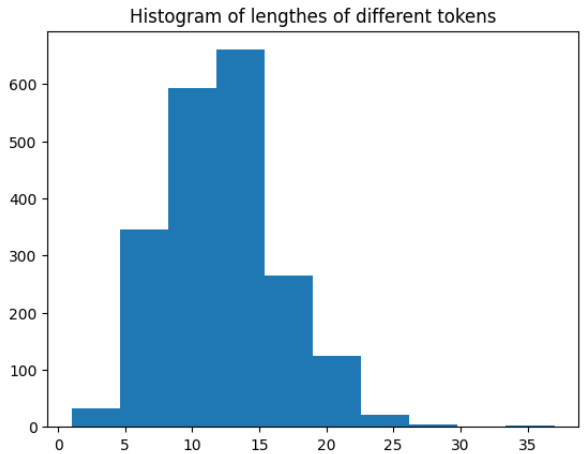
تنوع token ها بسیار گسترده است و طبق شکل زیر ۲۸۴ نوع از token داریم و نشان دادن خودِ آن‌ها در این گزارش دشوار است؛ لذا آن‌ها در فایل note book موجود هستند.

```
We have  284  Data, so we should consider 284 One-hot vectors
The longest smile has 37 characters
```

شکل ۱-۳- تنوع token ها

```
The frequency of the tokens is
 [1.0386e+04 2.5000e+01 3.6100e+02 1.4500e+02 4.3000e+01 6.0000e+00
 3.0000e+00 1.0000e+00 2.0000e+00 4.0000e+00 4.0000e+00 4.2000e+01
 2.1000e+01 7.0000e+00 3.0000e+00 1.0000e+00 1.0000e+00 6.0000e+00
 2.0000e+00 2.0000e+00 3.0000e+00 2.0000e+00 1.0491e+04 5.9000e+01
 7.0000e+00 1.0190e+03 4.0000e+00 2.5500e+02 8.3000e+01 3.3000e+01
 4.0000e+00 4.0000e+00 6.8000e+01 2.2000e+01 5.0000e+00 1.0000e+00
 5.0000e+00 1.0000e+00 7.0000e+00 2.0000e+00 3.0000e+00 1.0000e+00
 2.0000e+00 1.0000e+00 1.8400e+02 2.0000e+00 2.3000e+01 8.0000e+00
 4.0550e+03 1.0000e+00 2.0000e+00 1.8000e+01 4.1800e+02 3.0600e+02
 2.6300e+02 1.2000e+02 7.8000e+01 4.0000e+00 3.0000e+00 4.0000e+00
 1.0000e+00 1.9000e+01 1.3000e+01 3.0000e+00 2.0000e+00 2.0000e+00
 1.0000e+00 6.0000e+00 4.8000e+01 3.1000e+01 1.0000e+01 3.0000e+00
 1.0000e+00 1.0000e+00 1.0000e+00 4.0000e+00 1.0000e+00 1.0000e+00
 4.0000e+00 1.0000e+00 5.0000e+00 5.0000e+00 7.0000e+00 1.0000e+00
 8.0000e+00 1.0000e+01 4.0000e+00 3.3080e+03 2.0000e+00 1.0000e+00
 9.0000e+00 3.0000e+00 3.3100e+02 1.0000e+00 1.8100e+02 1.4500e+02
 1.0700e+02 8.8000e+01 7.0000e+00 2.0000e+00 3.0000e+00 1.0000e+00
 6.0000e+00 1.0000e+00 8.0000e+00 2.0000e+00 4.1000e+01 2.4000e+01
 8.0000e+00 3.0000e+00 2.0000e+00 6.0000e+00 1.0000e+00 1.0000e+00
 2.0000e+00 1.0000e+00 2.5130e+03 1.0000e+00 5.0000e+00 3.0900e+02
 2.3600e+02 2.1000e+02 1.6400e+02 1.5100e+02 4.0000e+00 1.0000e+00
 2.0000e+00 1.0000e+00 1.0000e+00 4.0000e+00 1.0000e+00 1.0000e+00
 1.0000e+00 1.0000e+00 8.0000e+00 1.0000e+00 2.3000e+01 1.0000e+01
 4.0000e+00 1.0000e+00 1.0000e+00 1.0000e+00 7.0000e+00 1.0000e+00
 5.0000e+00 1.0000e+00 9.0000e+00 6.0000e+00 8.0000e+00 4.0000e+00
 2.0000e+00 1.0000e+00 1.4000e+03 6.0000e+00 1.0000e+00 1.9600e+02
 1.2400e+02 1.0500e+02 6.4000e+01 6.0000e+01 4.0000e+00 2.0000e+00
 1.0000e+00 1.0000e+00 1.0000e+00 1.0000e+01 1.8000e+01 4.0000e+00
 1.0000e+00 1.0000e+00 4.0000e+00 2.0000e+00 1.0000e+00 4.4800e+02
 2.0000e+00 5.1000e+01 4.0000e+01 3.4000e+01 1.5000e+01 1.4000e+01
 1.0000e+00 4.0000e+00 2.0000e+00 1.0000e+00 1.0000e+00 1.1600e+02
 1.0000e+00 1.8000e+01 1.3000e+01 8.0000e+00 6.0000e+00 5.0000e+00
 1.0000e+00 1.0000e+00 1.0000e+00 5.4000e+01 9.0000e+00 7.0000e+00
 4.0000e+00 2.0000e+00 2.5000e+01 6.0000e+00 2.0000e+00 1.0000e+00
 1.6000e+01 1.0000e+00 1.0000e+00 8.0340e+03 5.0000e+01 5.1000e+01
 6.0000e+00 1.0000e+00 1.0000e+00 2.8983e+04 5.0000e+01 3.0000e+00
 2.0000e+00 2.0000e+01 1.2430e+03 1.0000e+00 1.3540e+03 4.0000e+00
 3.2300e+02 7.2000e+01 2.0000e+00 1.0000e+00 2.3700e+02 6.1800e+02
 4.3000e+01 5.0000e+00 1.6000e+01 2.4000e+01 4.9100e+02 8.0000e+00
 1.4500e+02 4.3000e+01 3.0000e+00 5.0000e+00 3.8700e+03 1.0000e+00
 4.9000e+01 4.0000e+00 1.0000e+00 8.0000e+00 1.4000e+01 1.4000e+01
 2.1000e+01 1.9000e+01 7.9000e+01 6.1140e+03 2.0000e+00 1.0000e+00
 6.4000e+01 3.0000e+01 4.0000e+00 1.3600e+02 1.1000e+01 5.0700e+02
 1.0000e+00 1.0000e+00 2.0000e+00 6.3000e+01 3.1490e+03 4.3000e+01
 2.0000e+00 1.0000e+00 1.4290e+03 2.0000e+00 3.0000e+00 8.1000e+01
 5.0000e+00 1.0000e+00 3.0000e+01 2.0000e+00 1.0000e+00 5.0000e+00
 3.0000e+00 5.0000e+00 1.0000e+00 1.0000e+00 1.0000e+00 0.0000e+00
 0.0000e+00 0.0000e+00]
```

شکل ۱-۴- فرکانس رخداد token ها



شکل ۱-۵- توزیع token ها؛ همانطور که در ابتدا برای شکل ۱-۱- حدس زده شد، توزیع tokenها به صورت یکنواخت نیست و لذا برای قرار دادن آن‌ها در ۱۰ تا bin مختلف، بهتر است به صورت غیر یکنواخت، آن‌ها را انتخاب کنیم.

قسمتهای ب، ج د

دقت شبکه‌های بازگشتی و شبکه MLP برای داده‌های Train و Test به صورت زیر هستند:

```
Accuracy for train data using MLP network is:  89.24278846153845
Accuracy for test data using MLP network is:  77.1819526627219
Accuracy for train data using LSTM network is:  89.9639423076923
Accuracy for test data using LSTM network is:  81.04659763313609
Accuracy for train data using BiLSTM network is:  86.17788461538461
Accuracy for test data using BiLSTM network is:  81.56434911242604
```

شکل ۲-۱- دقت شبکه‌های MLP، LSTM و BiLSTM برای داده‌های Train و Test

قسمت ه

دقت شبکه‌های بازگشتی و MLP برای 5-Cross Validation به صورت زیر هستند:

```
Best Accuracy for test data using MLP network is:  83.06213017751479 %
Average Accuracy for test data using MLP network is:  70.1405325443787 %
Best Accuracy for test data using LSTM network is:  82.8957100591716 %
Average Accuracy for test data using LSTM network is:  72.31878698224851 %
```

شکل ۳-۱- دقت شبکه‌های LSTM و MLP با استفاده از 5-Cross Validation

قسمت و

فرمولهای شیمیایی مطرح شده، هم از ابتدا به انتها وابسته هستند و هم از انتها به ابتدا.
لذا شبکه های BiLSTM میتوانند عملکرد طبقه بندی آنها را بهبود ببخشند؛ زیرا این شبکه ها با در نظر گرفتن هر داده به صورت One-Hot، یک دور توالی فرمول شیمیایی را از ابتدا تا انتهای آن بررسی میکنند و یک بار نیز از انتها تا ابتدا بررسی میکنند.

قسمت ز

دقت شبکه‌های بازگشتی و MLP برای در نظر گرفتن ۱۰ عدد bin به صورت زیر هستند:

```
Accuracy for test data for different bins using MLP network is:
  [87.08333333333333, 88.54166666666666, 80.625, 80.0, 84.375, 63.21022727272727, 79.16666666666666, 69.23076923076923, 71.42857142857143, 33.33333333333333]
Accuracy for test data for different bins using BiLSTM network is:
  [82.6086956521739, 86.25, 83.78378378378379, 88.0, 79.22077922077922, 69.76744186046511, 87.5, 61.53846153846154, 47.61904761904761, 47.61904761904761]


The First accuracy for MLP is: 87.08333333333333
The First accuracy for BiLSTM is: 82.6086956521739


The Last accuracy for MLP is: 33.33333333333333
The Last accuracy for BiLSTM is: 47.61904761904761
```

شکل ۴-۱- دقت شبکه‌های BiLSTM و MLP برای bin های مختلف.

میدانیم که در یک شبکه Sequential، ارتباط مکانی(یا زمانی) نمونه ها بررسی میشود. لذا از این جهت، نسبت به شبکه های MLP برتری دارند که این اتفاق را برای بین های نهایی که طول آن‌ها بیشتر است، شاهد هستیم.

بنابراین اگر طول ورودی بسیار زیاد باشد، شبکه های LSTM و یا BiLSTM میتوانند بهتر از MLP عمل کنند. اما هنگامی که طول ورودی کم است، شبکه هایMLP بهتر عمل خواهند کرد. زیرا شبکه های LSTM، در این صورت ممکن است دچار اوورفیتینگ شوند؛ چون داده‌ی موجود، برای بین های اول، بسیار کم حجم است و نیازی به استفاده از یک شبکهLSTM برای ترین کردن آن نمیباشد و استفاده از LSTM به Performance شبکه نمیتواند کمک خاصی کند.