

Data Communication

Babol Noshirvani University of Technology (NIT)

Final Project

Public Wi-Fi Latency Analysis

یک فایل داده با نام wifi_latency.csv در اختیار شما قرار می‌گیرد که از ثبت ترافیک یک شبکه‌ی وای-فای عمومی تهیه شده است. هر ردیف نماینده‌ی یک جریان (flow) دو دقیقه‌ای است. متغیرهای اصلی عبارت‌اند از:

Column Name	Description	DataType
latency_ms	میانگین تأخیر رفت-و-برگشت (RTT) بر حسب میلی‌ثانیه	Numeric
rss_i_dbm	توان سیگنال دریافت شده (dBm-)	Numeric
snr_db	نسبت سیگنال به نویز	Numeric
band	باند فرکانسی اتصال (2.4 GHz یا 5 GHz)	Categorical
channel_util%	درصد اشغال کانال در زمان نمونه‌گیری	Numeric
num_assoc_devices	تعداد دستگاه‌های متصل به همان اکسس‌پوینت	Numeric
client_speed_mbps	حداکثر نرخ فیزیکی لینک گزارش‌شده توسط مشتری	Numeric
ap_vendor	سازنده‌ی اکسس‌پوینت (Cisco, Aruba, TP-Link, ...)	Categorical
protocol	نوع پروتکل 802.11 (n, ac, ax)	Categorical
distance_m	فاصله تقریبی کلاینت تا اکسس‌پوینت	Numeric

هدف اصلی این پروژه تعیین این است که کدام متغیرها بیشترین نقش را در افزایش یا کاهش تأخیر دارند. برای این کار، از Mutual Information بین latency_ms و سایر ویژگی‌ها استفاده می‌کنیم.

گام‌های انجام پروژه:

1 . Data Preparation

- دیتاست مربوطه را از این لینک دانلود و به پروژه خود اضافه کنید.

2 . Data Cleaning

- مقادیر تهی، تکراری، outliers و تبدیل تایپ (در صورت نیاز) را با روش دلخواه مدیریت کنید.

3 . Manipulation and Transformation

- انتخاب/حذف ستون‌ها، فیلتر کردن سطرها، ایجاد ستون‌های جدید (مثلاً طبقه‌بندی rssi_dbm به خوب/متوسط/ضعیف).
- نرمال‌سازی یا مقیاس‌بندی ویژگی‌های عددی.
- در صورت لزوم ارزیابی هم‌بستگی و حذف ویژگی‌های بسیار هم‌بسته.

4 . Core Task

- برای هر ویژگی ۶ بازه/گروه تعریف کنید (اگر عددی است؛ برای اسمی همان مقادیر را نگه دارید).
- احتمال‌های $P(x)$, $P(y)$ و $P(x,y)$ را طبق فرمول‌های درس انتقال داده محاسبه کنید.
- مقدار MI بین latency_ms و هر ویژگی را به دست آورید.
- جدول نتایج را بر اساس MI نزولی مرتب کنید و مهم‌ترین عوامل را گزارش دهید.

5 . Data Visualization

- حداقل سه نمودار گویا (هیستوگرام توزیع تأخیر، نمودار جعبه‌ای تأخیر بر حسب باند، نقشه حرارتی MI).
- فقط از کتابخانه‌های مجاز (Plotly, Matplotlib). استفاده از Scikit-learn یا کتابخانه‌های Machine Learning ممنوع است.

6 . Documentation

- مراحل، فرضیات، و یافته‌های کلیدی خود را به صورت کامل در یک فایل PDF توضیح دهید.
- بیان کنید که کدام عوامل (به‌طور مثال channel_util % یا band) بیشترین سهم را در افزایش تأخیر داشته‌اند و چرا این نتیجه منطقی است.
- هر پیشنهاد بهبود پروژه را به گزارش خود اضافه کنید.

نکات تکمیلی پروژه

- قبل از اعمال MI، اگر latency_ms توزیع دُم-کلفت داشت، می‌توانید آن را در مقیاس لگاریتمی بازه‌بندی کنید.
- برای ویژگی‌های عددی، یک تابع کمکی بنویسید که به‌صورت خودکار حداقل (min) و حداکثر (max) را خوانده و آن را به ۶ بازه مساوی تقسیم کند.
- اگر ستون distance_m وجود نداشت یا مقداردهی نشده بود، با میانگین فاصله تقریباً 15 m جایگزین کنید.
- همه محاسبات احتمال را روی کل نمونه تقسیم بر N انجام دهید.

فرمول‌های مرجع درس

1. :

$$P(x) = \frac{\text{count}(x)}{N} \quad (1)$$

2. :

$$P(x, y) = \frac{\text{count}(x \cap y)}{N} \quad (2)$$

3. :

$$I(X; Y) = \sum_x \sum_y P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)} \quad (3)$$

تحویل نهایی

- کد Python (یا فایل Excel دارای فرمول‌ها)
- فایل گزارش مستند (PDF)
- سه تصویر نمودار خروجی
- فایل wifi_latency.csv پردازش‌شده (پس از حذف ستون‌های غیرضروری)

آپلود موارد ذکر شده به صورت آنلاین در کوئرا و ارائه پروژه به صورت حضوری خواهد بود.
زمان و مکان ارائه نیز متعاقبا اعلام خواهد شد.