تمرین چهارم

مبانی علوم داده – دکتر نادری

1. Regularization چطور کار می‌کند؟ چگونه باعث جلوگیری از Overfitting میشود؟
2. کدام نوع Regularization برای Feature Selection مناسب است؟ L1 یا L2؟ چگونه؟
3. تفاوت Domain Adaptation و Transfer Learning چیست؟ مقایسه کنید.
4. چطور K را در الگوریتم K-means انتخاب می‌کنیم؟ چگونه ممکن است در این الگوریتم Overfitting رخ دهد؟ توضیح دهید.
5. Bias-Variance Tradeoff را ابتدا به صورت ساده توضیح دهید. سپس یک نمونه از این مقایسه با کمک MAE, MSE برای بررسی مدل مثال بزنید و محدودیت های هر Metric را بررسی کنید.
6. فرض کنید یک دیتاست داریم و می‌خواهیم یک مدل ماشین لرنینگ روی آن آموزش بدهیم، آیا برای افزایش دقت می‌توانیم پیچیدگی مدل را هر چقدر خواستیم زیاد کنیم؟ برای دستیابی به حداکثر دقت باید به چه عواملی توجه کنیم؟ توضیح دهید.
7. Precision/Recall tradeoff را با یک مثال توضیح دهید و Confusion matrix مربوطه را بنویسید.
8. مسئله Precision-Recall Tradeoff:
   - یک نفر فقط از 5 مدل کفش خوشش می‌آید و به کفش فروشی می‌رود. فروشنده به خریدار 100 مدل کفش پیشنهاد می‌دهد و بین این 100 مدل، 4 مدل از آن دسته ای است که خریدار خوشش می‌آید.

Precision و Recall فروشنده چقدر است؟ Confusion Matrix را بسازید.

- حالت دوم: فروشنده 4 مدل پیشنهاد می‌دهد و خریدار از 3 مدل خوشش می‌آید. در این حالت Precision و Recall چقدر است؟ برای این مثال هم Confusion Matrix را بنویسید.

9. یک مدل برای تشخیص Fraud در سیستم های مالی توسعه داده‌ایم. نسبت حضور هر کلاهبرداری به انتقال سالم 1 به 10000 است دیتاست هم از همین نسبت پیروی می‌کند. توضیح دهید چگونه این سیستم را آموزش می‌دهید و ارزیابی می‌کنید؟
10. مفهوم منحنی Receiver Operating Characteristic (ROC) و Area Under the Curve (AUC)را توضیح دهید.

- چگونه امتیاز AUC را در مسائل طبقه‌بندی باینری تفسیر می‌کنید؟ اگر AUC یک مدل 0.7 باشد در مقابل 0.9 چه معنایی دارد؟
- تفاوت Precision-Recall curve را با ROC بررسی کنید ومثالی بزنید که Precision-Recall curve برای ارزیابی مناسب تر باشد.

11. فرآیند K-fold cross-validation را توضیح دهید و چگونه برای ارزیابی عملکرد مدل استفاده می‌شود. مزایا و معایب K-fold cross-validation در مقایسه با تقسیم ساده داده‌ها به مجموعه آموزش و آزمایش چیست؟ چگونه مقدار k را انتخاب می‌کنید؟
12. فرض کنید به خواسته saman-kala.ir یک سیستم پیشنهاد دهنده کالا برای مشتریان طراحی کرده‌اید. در کمال تعجب سیستم پیشنهاد دهنده‌ی قبلی، به صورت دستی توسط کارمندان انجام می‌شد!

- حال مدیر پروژه از شما می‌خواهد پس از اجرای پروژه پیشنهاد دهنده خودتان، به آن‌ها نشان دهید مدل شما از مدل قبلی بهتر است.(راهنمایی: برای اینکه بگوییم یک الگوریتم

بهتر است باید یک (یا چند) معیار داشته باشیم که بتوانیم مقایسه کنیم. شما به عنوان دیتاساینتست باید در تعریف این معیار ها کمک کنید.)

- همچنین مدیر سایت، از تیم دیتا، تخمین زمان میخواهد که هزینه این پروژه چه مدت پس از اجرا شدن آن، توسط سودش برمی‌گردد (ROI).پس باید روشی را پیشنهاد دهید که مشخص کند این پروژه چقدر برای شرکت سودآوری دارد.

  (راه دستیابی به این جواب ها را با ذکر فرض‌های منطقی خودتان، توضیح دهید.)

موفق باشید.