**علی قنبری دمشهری**

**درس: مباحث ویژه**

**استاد: جناب محمد احمدزاده**

***سؤالات*:**

**1:** **تفاوت Supervised و Unsupervised Learning**

**در Supervised Learning، داده‌ها شامل برچسب (Label) هستند و مدل از روی ورودی‌ها و خروجی‌های مشخص، الگو را یاد می‌گیرد.**

**در مقابل، درUnsupervised Learning داده‌ها بدون برچسب‌اند و مدل باید خود، ساختار یا الگوهای پنهان را کشف کند، (مثل خوشه‌بندی یا کاهش ابعاد).**

**2:** **ضرورت Feature Scaling در الگوریتم‌های Machine Learning**

**الگوریتم‌هایی مانند SVM، KNN و Gradient Descent به مقیاس داده‌ها حساس هستند. اگر ویژگی‌ها در مقیاس‌های متفاوت باشند، نتایج یادگیری منحرف می‌شوند Feature Scaling کمک می‌کند تا همه ویژگی‌ها در محدوده مشابهی قرار گیرند و مدل عملکرد بهتری داشته باشد.**

**3:** **تفاوت Standardization و Normalization**

**Normalization داده‌ها را به بازه [0,1] می‌برد اغلب با Min-Max در حالی که Standardization داده‌ها را با میانگین صفر و انحراف معیار یک نرمال می‌کند. استفاده از هر کدام به نوع الگوریتم و توزیع داده‌ها بستگی دارد.**

**4:** **کاربرد Min-Max Normalization**

**Min-Max Normalization زمانی مفید است که داده‌ها به‌طور یکنواخت در یک بازه خاص (مثل [0,1]) نیاز به مقایسه یا پردازش دارند، مخصوصاً برای الگوریتم‌هایی که به فاصله‌های عددی حساس‌اند ، مانند KNN یا شبکه‌های عصبی.**

**5:** **E. Z-Score Normalization چیست و چرا استفاده می‌شود؟**

**در Z-Score، هر مقدار از میانگین کم و بر انحراف معیار تقسیم می‌شود. این روش برای داده‌هایی که توزیع نرمال دارند و نسبت به مقیاس حساس هستند مانند PCA که مفید است.**

**6: F. Regularization در الگوریتم‌های Machine Learning چیست؟**

**Regularization تکنیکی برای جلوگیری از Overfitting است. با افزودن یک جریمه (penalty) به تابع هزینه مثلاً L1 یا L2، پارامترهای مدل محدود می‌شوند تا از پیچیدگی بیش از حد جلوگیری شود.**

**7: G. Overfitting و Underfitting چه مشکلاتی در Model- building ایجاد می‌کنند؟**

**Overfitting زمانی رخ می‌دهد که مدل بیش از حد با داده‌های آموزش سازگار شود و در داده‌های جدید عملکرد ضعیفی داشته باشد. Underfitting زمانی است که مدل قادر به یادگیری الگوها حتی از داده‌های آموزشی نیز نیست. هر دو باعث کاهش دقت و تعمیم‌پذیری مدل می‌شوند.**

**8:** **کاربرد Train/Test Split و Cross-Validation**

**Train/Test Split برای ارزیابی سریع مدل با تقسیم داده‌ها به بخش‌های آموزش و آزمون است. اما Cross-Validation مثلاً K-Fold روشی دقیق‌تر برای ارزیابی عملکرد مدل است که به تعمیم‌پذیری بیشتر کمک می‌کند و به کاهش بایوس کمک می‌نماید.**

**9: Gradient Descent چگونه کار می‌کند؟**

**Gradient Descent یک الگوریتم بهینه‌سازی است که با استفاده از گرادیان (مشتق) تابع هزینه، پارامترهای مدل را در جهت کاهش خطا به‌روزرسانی می‌کند. با تکرار این فرایند، مدل به مقدار مینیمم تابع هزینه نزدیک می‌شود.**

**10: چرا Deep Learning برای پیچیده‌ترین مسائل استفاده می‌شود؟**

**Deep Learning با استفاده از لایه‌های عمیق شبکه‌های عصبی قادر است ویژگی‌های پیچیده و انتزاعی را از داده‌ها استخراج کند. این توانایی آن را برای پردازش زبان طبیعی، بینایی ماشین و مسائل پیچیده مانند تشخیص چهره یا ترجمه خودکار ایده‌آل می‌سازد.**

**11:** **چرا Python زبان برنامه‌نویسی محبوب علم داده است؟**

**Python به دلیل سادگی سینتکس، پشتیبانی گسترده از کتابخانه‌های علمی مانند NumPy، Pandas، Matplotlib، Scikit-learn و جامعه کاربری فعال، انتخاب اول در حوزه علم داده محسوب می‌شود.**

**12: B. Pandas و NumPy چه تفاوتی دارند؟**

**NumPy بیشتر برای محاسبات عددی با آرایه‌های چندبعدی استفاده می‌شود، در حالی‌که Pandas برای تحلیل و پردازش داده‌های ساختاریافته مانند جدول‌ها (DataFrame) طراحی شده است.**

**13:** **چرا Matplotlib برای تجسم داده‌ها استفاده می‌شود؟**

**Matplotlib یک کتابخانه پایه و انعطاف‌پذیر برای تولید نمودارهای دو بعدی است که کنترل کامل بر روی اجزای نمودار (محورها، برچسب‌ها، رنگ‌ها) را فراهم می‌سازد.**

**14:** **چرا Seaborn برای تجسم داده‌های پیشرفته کاربرد دارد؟**

**Seaborn بر پایه Matplotlib ساخته شده و امکانات تجسم آماری پیشرفته‌تری مانند نمودارهای جفتی، جعبه‌ای و گرمایی را با ظاهری زیباتر و کدنویسی ساده‌تر فراهم می‌کند.**

**15: چگونه می‌توانید یک Function در Python تعریف کنید؟**

**با استفاده از کلمه کلیدی def، مانند:**

**def greet(name):**

**return f"Hello, {name}"**

**16:** **چرا از List Comprehension در Python استفاده می‌شود؟**

**List Comprehension روشی خلاصه و بهینه برای ایجاد لیست‌ها با استفاده از حلقه‌ها و شروط در یک خط کد است. موجب افزایش خوانایی و کاهش حجم کد می‌شود.**

**17: چگونه می‌توان یک CSV file را در Python خواند؟**

**با استفاده از کتابخانه Pandas**

**import pandas as pd**

**df = pd.read\_csv("filename.csv")**

**18: H. JSON و XML چه تفاوتی دارند؟**

**JSON ساختاری ساده، خواناتر و مناسب‌تر برای داده‌های سبک است و بیشتر در APIها استفاده می‌شود. XML ساختاری سنگین‌تر و توصیفی‌تر دارد و برای اسناد پیچیده‌تر کاربرد دارد.**

**19: List و Array چه تفاوتی دارند؟**

**در Python، list ساختاری داینامیک با قابلیت نگهداری انواع داده‌ مختلف است، در حالی که ماژول array فقط مقادیر یک نوع داده را نگه می‌دارد و برای کاربردهای عددی بهینه‌تر است.**

**20: Dictionary در Python چگونه کار می‌کند؟**

**Dictionary ساختاری کلید-مقدار (key-value) است که از Hashing برای دسترسی سریع به مقادیر استفاده می‌کند.**

**21: Tuple و List چه تفاوتی دارند؟**

**Tuple غیرقابل‌تغییر (immutable) است، یعنی پس از تعریف نمی‌توان عناصر آن را تغییر داد، در حالی که List قابل‌تغییر است (mutable). بنابراین Tuple برای داده‌های ثابت ایمن‌تر و سریع‌تر است.**

**22:** **چرا در Python از Set برای حذف داده‌های تکراری استفاده می‌شود؟Set مجموعه‌ای بدون ترتیب و تکرار است. هنگام اضافه‌کردن عناصر تکراری، آن‌ها را به طور خودکار حذف می‌کند و از لحاظ عملکردی سریع‌تر از List در این زمینه است.**

**23: Stack و Queue چه تفاوتی دارند؟**

**Stack از اصل LIFO (آخر وارد، اول خارج) و Queue از اصل FIFO (اول وارد، اول خارج) پیروی می‌کنند. کاربردها و عملکرد آن‌ها در مدیریت داده‌ها متفاوت است.**

**24: Hash Table چیست و چرا کاربرد دارد؟**

**ساختاری برای نگهداری داده‌ها به صورت کلید-مقدار است که با استفاده از تابع هش، دسترسی بسیار به داده‌ها را فراهم می‌کند. در ساختارهایی مثل Dictionary و Set استفاده می‌شود.**

**25: Binary Tree و B-Tree چه تفاوتی دارند؟**

**Binary Tree حداکثر دو فرزند برای هر گره دارد و بیشتر در حافظه اصلی استفاده می‌شود، در حالی که B-Tree چندین فرزند دارد و برای ذخیره‌سازی داده‌ها در پایگاه داده‌ها و دیسک بهینه شده است.**

**26:** **چرا Graph Data Structure برای شبکه‌های اجتماعی استفاده می‌شود؟**

**شبکه‌های اجتماعی بر پایه گره‌ها (افراد) و یال‌ها (روابط) هستند، که ساختار گراف برای مدل‌سازی چنین ارتباطات پیچیده‌ای مناسب‌ترین انتخاب است.**

**27:** **چرا Dynamic Programming در حل مسائل پیچیده کاربرد دارد؟**

**Dynamic Programming زیرمسائل تکراری را با استفاده از حافظه‌سازی memoization یا tabulation حل می‌کند، و در مسائل بهینه‌سازی با ساختار بازگشتی بسیار مؤثر است.**

**28: Recursion چیست و چرا در الگوریتم‌های پیشرفته استفاده می‌شود؟**

**Recursion به معنی فراخوانی تابع توسط خودش است و در حل مسائل بازگشتی مانند پیمایش درخت‌ها، گراف‌ها و الگوریتم‌های تقسیم‌و‌حل (Divide & Conquer) بسیار کاربردی است.**

**29: چرا Data Cleaning در علم داده اهمیت دارد؟  
زیرا داده‌های خام اغلب دارای نویز، خطا و نواقص هستند.**

**پاک‌سازی داده‌ها کیفیت ورودی مدل را افزایش می‌دهد و منجر به نتایج دقیق‌تر می‌شود.**

**30: Missing Values چگونه مدیریت می‌شوند؟  
از روش‌هایی نظیر حذف سطر/ستون دارای مقادیر گمشده، میانگین‌گیری، مدلسازی داده‌های گمشده یا استفاده از الگوریتم‌هایی مقاوم در برابر داده ناقص استفاده می‌شود.**

**31: Outliers چیست و چگونه می‌توانید آن‌ها را تشخیص دهید؟  
مقادیر پورت داده‌هایی هستند که به‌طور قابل‌توجهی با سایر مقادیر تفاوت دارند. می‌توان آن‌ها را با روش‌هایی مانند Z-Score، IQR یا نمودار Boxplot تشخیص داد.**

**32: چرا Data Transformation کاربرد دارد؟  
برای تغییر شکل داده‌ها به فرم مناسب جهت تحلیل، افزایش دقت مدل، نرمال‌سازی توزیع و کاهش پیچیدگی داده‌ها استفاده می‌شود.**

**33: Encoding Techniques (One-Hot Encoding و Label Encoding) چه تفاوتی دارند؟**

**Label Encoding مقادیر متنی را به اعداد ترتیبی تبدیل می‌کند، در حالی که One-Hot Encoding برای هر مقدار یک بُعد جداگانه ایجاد می‌کند،**

**که از ورود ترتیب کاذب جلوگیری می‌کند.**

**34:** **چرا Feature Selection در Model-building اهمیت دارد؟**

**باعث حذف ویژگی‌های بی‌اهمیت، افزایش دقت مدل، کاهش پیچیدگی محاسبات و جلوگیری از Overfitting می‌شود.**

**35: Duplicate Data چگونه در پایگاه داده‌ها حذف می‌شود؟**

**با استفاده از توابع شناسایی تکرارها drop\_duplicates در pandas**

**بر اساس شناسه‌ها یا ترکیبی از ویژگی‌ها.**

**36: Irrelevant Data چه مشکلاتی را در پیش‌بینی‌های Machine Learning ایجاد می‌کند؟**

**باعث کاهش دقت مدل، افزایش نویز، پیچیدگی بیشتر و طولانی شدن زمان آموزش می‌شود.**

**37: چرا Data Imputation برای پر کردن Missing Values کاربرد دارد؟برای جلوگیری از حذف داده‌های باارزش و حفظ ساختار داده، با استفاده از میانگین، مد، مدل‌های پیش‌بینی یا تکنیک‌های آماری داده‌های گمشده پر می‌شوند.**

**38: چگونه می‌توانید Normality را در داده‌های عددی بررسی کنید؟**

**با استفاده از نمودار هیستوگرام، نمودار Q-Q و آزمون‌های آماری مانند Shapiro-Wilk یا Kolmogorov-Smirnov**

**39: چرا برای نمایش رابطه‌های خطی از Line Chart استفاده می‌شود؟  
زیرا Line Chart امکان نمایش روند تغییرات یک متغیر پیوسته در طول زمان یا نسبت به متغیر دیگر را به‌صورت دقیق و قابل درک فراهم می‌کند.**

**40: چرا برای مقایسه داده‌های گروهی از Bar Chart استفاده می‌شود؟  
Bar Chart برای نمایش و مقایسه اندازه یا فراوانی دسته‌های مجزای داده‌ها کاربرد دارد و تفاوت بین گروه‌ها را به‌خوبی نشان می‌دهد.**

**41: چرا برای نمایش رابطه‌های غیرخطی از Scatter Plot استفاده می‌شود؟  
Scatter Plot امکان نمایش پراکندگی و نوع رابطه بین دو متغیر را فراهم می‌کند و الگوهای غیرخطی یا خوشه‌ها را آشکار می‌سازد.**

**42: چرا برای نمایش سه متغیر از Bubble Chart استفاده می‌شود؟  
Bubble Chart با استفاده از محور X، محور Y و اندازه حباب، امکان نمایش همزمان سه ویژگی عددی را فراهم می‌کند.**

**43: چرا برای نمایش رابطه‌های بین متغیرها از Heatmap استفاده می‌شود؟  
Heatmap از شدت رنگ برای نمایش ماتریس همبستگی یا روابط بین متغیرها استفاده می‌کند و الگوهای کلی را به‌صورت بصری نمایان می‌سازد.**

**44: چرا برای تحلیل روابط بین متغیرها از Pairplot استفاده می‌شود؟  
Pairplot مجموعه‌ای از Scatter Plotها و Histogramها برای تمام جفت‌های متغیرها تولید می‌کند که به تحلیل همزمان چند رابطه کمک می‌کند.**

**45: چرا برای تشخیص Outliers از Boxplot استفاده می‌شود؟  
Boxplot با نمایش چارک‌ها و حد آستانه، مقادیر پرت را به‌وضوح مشخص کرده و در تشخیص داده‌های غیرعادی بسیار مؤثر است.**

**46: چرا برای نمایش توزیع داده‌ها از Histogram استفاده می‌شود؟  
Histogram داده‌ها را به بازه‌های پیوسته تقسیم کرده و فراوانی هر بازه را نمایش می‌دهد و به تحلیل چگالی توزیع کمک می‌کند.**

**47: چگونه می‌توانید یک 3D Plot در Python ایجاد کنید؟  
با استفاده از کتابخانه matplotlib و ماژول Axes3D می‌توان نمودارهای سه‌بعدی مانند سطحی یا پراکندگی سه‌بعدی رسم کرد.**

**48: چرا Seaborn برای تجسم داده‌های پیشرفته استفاده می‌شود؟  
زیرا Seaborn ابزارهای سطح بالا برای تجسم آماری، زیبایی بصری بالا، و قابلیت تلفیق با pandas و matplotlib را داراست.**

**49: چرا Distributed Ledger Technology (DLT) در مدیریت داده‌ها کاربرد دارد؟  
DLT با فراهم‌سازی ساختاری غیرمتمرکز، شفاف، و تغییرناپذیر برای ذخیره تراکنش‌ها، در تضمین صحت داده‌ها و کاهش وابستگی به واسطه‌ها نقش کلیدی دارد.**

**50: چرا Blockchain برای ذخیره‌سازی داده‌های امن استفاده می‌شود؟  
Blockchain با بهره‌گیری از رمزنگاری، اجماع توزیع‌شده و ساختار زنجیره‌ای، تضمین می‌کند که داده‌ها تغییرناپذیر، قابل رهگیری و مقاوم در برابر نفوذ باشند.**

**51: چرا GANs (Generative Adversarial Networks) در علم داده پیشرفته کاربرد دارند؟  
GANها قادرند داده‌های مصنوعی با کیفیت بالا تولید کنند و در زمینه‌هایی مانند افزایش داده، بازسازی تصاویر، و تولید محتوای جدید بسیار مؤثرند.**

**52: D. PCA و t-SNE چه تفاوتی دارند؟  
PCA یک روش خطی برای کاهش ابعاد است که به حفظ واریانس کمک می‌کند، در حالی‌که t-SNE روشی غیرخطی است که برای نمایش بصری خوشه‌ها در داده‌های پیچیده مناسب‌تر است.**

**53: چرا UMAP برای Dimensionality Reduction استفاده می‌شود؟  
UMAP هم ساختار جهانی و هم ساختار محلی داده‌ها را حفظ کرده و نسبت به t-SNE سریع‌تر و مقیاس‌پذیرتر است، که آن را برای تحلیل داده‌های بزرگ مناسب می‌سازد.**

**54: چرا NetworkX برای تحلیل گراف‌ها کاربرد دارد؟  
NetworkX مجموعه‌ای جامع از ابزارها برای ساخت، تحلیل و تجسم گراف‌ها فراهم می‌کند که در شبکه‌های اجتماعی، بیوانفورماتیک و تحلیل ارتباطات بسیار مفید است.**

**55: چرا NLTK برای پردازش زبان طبیعی (NLP) استفاده می‌شود؟  
NLTK کتابخانه‌ای قدرتمند در پایتون برای انجام وظایف زبانی مانند توکن‌سازی، برچسب‌گذاری نحوی، ریشه‌یابی و تحلیل معناشناسی است که در تحقیقات و آموزش NLP به‌کار می‌رود.**

**56:** **چرا Scrapy برای Web Scraping کاربرد دارد؟**

**Scrapy فریم‌ورکی سریع و مقیاس‌پذیر برای استخراج داده از وب‌سایت‌هاست که از قابلیت کراولینگ و پردازش موازی پشتیبانی می‌کند.**

**57:** **چرا BeautifulSoup برای کاری استفاده می‌شود؟**

**BeautifulSoup کتابخانه‌ای سبک و ساده برای استخراج اطلاعات از HTML و XML است که برای تحلیل سریع صفحات وب و پروژه‌های کوچک وب‌اسکرپینگ بسیار مناسب است.**

**The end**