سوال اول:

به عنوان مثال مسیله حدس زدن اعداد دست نوشته، یک مسیله دسته بندی است.

مسایل رگرسیون:

- برای پیش‌بینی میزان ثروت افراد
- در صنعت دارویی برای تست خون فوری
- در صنعت هتل‌داری برای پیش‌بینی ظرفیت مورد نیاز
- در صنعت بازی و پیدا کردن تبلیغ مناسب
- در ورزش برای پیش‌بینی تاثیر تمرینات مختلف

مسایل دسته بندی:

- Spam filtering
- Image classification
- Malware classification
- fraud detection
- Document classification

سوال دوم:

- Accuracy: درصد تاپل‌های مجموعه تست که به درستی طبقه بندی شده‌اند.

$$\textbf{Accuracy} = (TP + TN)/All$$

- Recall: کامل بودن - درصدی از تاپل‌های مثبت که طبقه‌بندی کننده به عنوان مثبت برچسب گذاری کرده است.

$$recall = \frac{TP}{TP + FN}$$

- Precision: دقت - درصدی از تاپل‌هایی که طبقه‌بندی کننده آنها را به عنوان مثبت برچسب گذاری کرده است و در واقع مثبت هستند.

$$precision = \frac{TP}{TP + FP}$$

- F1-Score: میانگین هارمونیک دو پارامتر بالا را می‌گویند.

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

سوال سوم:

فرض می‌کنیم مسیله پیدا کردن داشتن بیماری عرقی است.

حال باید برای ویژگی بیماری قلبی انتروپی را محاسبه کنیم:

$$Entropy(D) = -0.5 \times \log 0.5 - 0.5 \times \log 0.5 = 1$$

در ادامه Gain شاخه عروق خونی بسته را بررسی می‌کنیم:

$$D_{No} = [2-, 0+] = -1 \times \log 1 = 0$$

$$D_{Yes} = [0-, 2+] = -1 \times \log 1 = 0$$

$$Gain_{Closed\ Blood\ Vessels} = Entropy(D) - 0.5 \times Entropy(D_{No}) - 0.5 \times Entropy(D_{Yes}) = 1$$

از آنجایی که Gain ویژگی عروق بسته برابر ۱ شده است پس می‌توان این ویژگی را در ریشه درخت قرار داد و دسته‌بندی را انجام داد.

سوال چهارم:

فرض می‌کنیم مسیله پیدا کردن دوست داشتن سریال کلاه قرمزی است.

حال باید برای ویژگی دوست داشتن سریال کلاه قرمزی انتروپی را محاسبه کنیم:

$$Entropy(D) = -\frac{5}{7} \times \log \frac{5}{7} - \frac{2}{7} \times \log \frac{2}{7} = 0.862$$

شاخه سن را به سه دسته تقسیم می‌کنیم:

- کمتر از ۱۸ سال
- بین ۱۸ تا ۳۸ سال
- بالای ۳۸ سال

حال Gain را به دست می‌آوریم:

$$D_{x<18} = [2-, 0+] = -1 \times \log 1 = 0$$

$$D_{18<x<38} = [0-, 3+] = -1 \times \log 1 = 0$$

$$D_{38<x} = [2-, 0+] = -1 \times \log 1 = 0$$

$$Gain_{Age} = Entropy(D) - \frac{2}{7}Entropy(D_{<18}) - \frac{3}{7}Entropy(D_{18<x<38}) - \frac{2}{7}Entropy(D_{>38}) = 0.862$$

از آنجایی که Gain ویژگی سن برابر ۰.۸۶۲ شده است پس می‌توان این ویژگی را در ریشه درخت قرار داد و دسته‌بندی را انجام داد.

سوال پنجم:

این ویژگی میزان عدم خلوص را در یک نود نشان می‌دهد. روش محاسبه آن در فرمول زیر آمده است:

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

زمانی این ویژگی ماکسیمم یا ۱ می‌شود که از هرکدام از کلاس‌ها به تعداد برابر داشته باشیم که نشان دهنده ارزنده بودن آن اطلاعات می‌شود.

زمانی که در یک نود فقط یک کلاس داشته باشیم آنگاه این ویژگی مینیمم می‌شود که نشان دنده کم ارزش بودن این اطلاعات می‌باشد.

آنتروپی نیز شبیه به این ویژگی کار می‌کند. البته جینی اطلاعات بهتری نسبت انتروپی می‌دهد.

سوال ششم:

بیش برازش زمانی اتفاق می‌فتد که مدل بجای یادگیری از مجموعه یادگیری شروع به حفظ کردن آن کند. این باعث می‌شود که زمانی مجموعه تست یا جدید به مدل داده می‌شود، نتواند نتیجه مطلوبی بگیرد.

در روش Regularization پیچیدگی مدل را کنترل می‌کنیم و اجازه نمی‌دهیم مدل بیش برازش کند.

روش بعدی Drop Out نام دارد که به صورت رندوم وزن‌های شبکه را از پروسه یادگیری خارج می‌کند. این عمل باعث می‌شود که بیش برازش اتفاق نیافتد.