$\boxed{Q1}$

RV → Random Variable

A → RV representing color of ball drawn by Alice

B → RV representing color of ball drawn by Bob

Alice goes first → We know $P(A)$

Bob goes second → We know $P(B|A)$

1) $P(A=r, B=b) = P(A=r)\,P(B=b \mid A=r)$

$$= \left(\frac{2}{5}\right)\left(\frac{3}{4}\right)$$

2 red balls

out of 5 balls

3 blue balls of the remaining 4

$$= \boxed{\frac{3}{10}}$$

2) $P(A=b \mid B=b) = \dfrac{P(B=b \mid A=b)\,P(A=b)}{P(B=b)}$ [Baye's]

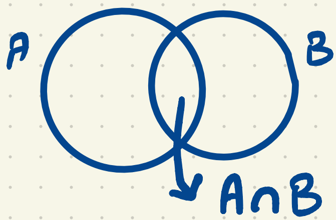
$$= \frac{P(B=b \mid A=b)P(A=b)}{P(B=b \mid A=b)P(A=b) + P(B=b \mid A=r)P(A=r)}$$

$$= \frac{(2/4)(3/5)}{(2/4)(3/5) + (3/4)(2/5)} = \boxed{\frac{1}{2}}$$

Q2 **a, c**

a) Inclusion-Exclusion Principle

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



b) To prove:

$$P(A \cup B) \leq P(A) + P(B) - P(A)P(B)$$

i.e. $P(A) + P(B) - P(A \cap B) \leq P(A) + P(B) - P(A)P(B)$

i.e. $$P(A)P(B) \leq P(A \cap B)$$

i.e. $$P(A)P(B) \leq P(A)P(B|A)$$

Not necessary. Proof by counter-example

$$P(A) = \frac{1}{2} \qquad P(B|A) = 1/3 \quad P(B|\bar{A}) = 2/3$$

$$\begin{aligned} P(B) &= P(B \cap A) + P(B \cap \bar{A}) \\ &= P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) \\ &= \frac{1}{3} \cdot \frac{1}{2} + \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{2} > P(B|A) \end{aligned}$$

$\boxed{Q2}$

c) True by Law of Total Probability

d) Since (c) is True

$$P(A) = P(A \cap C) + P(A \cap \bar{C})$$
$$= P(A|C)P(C) + P(A|\bar{C})P(\bar{C})$$

$\boxed{Q3}$ $\boxed{a, b, d}$

$\boxed{Q4}$ $\boxed{b, d}$

$\boxed{Q5}$ $\boxed{a, b, c, d}$

5a is non-trivial to prove

→ $X+W \sim$ Gaussian if X & W are independent Gaussians

→ $X+W \sim$ Gaussian if X & W are jointly Gaussian

→ $X-2Y \sim$ Gaussian by change of variable $W \to -2Y$ in previous result

Q6

$E \rightarrow$ Event that an email is spam

$S \rightarrow$ Event that an email is marked as spam

$P(E) = 0.01$

$P(S|E) = 0.9$

$P(S|\bar{E}) = 0.1$

$$P(E|S) = \frac{P(S|E)\,P(E)}{P(S)}$$

$$= \frac{P(S|E)\,P(E)}{P(S|E)P(E) + P(S|\bar{E})P(\bar{E})}$$

$$= \frac{(0.9)(0.01)}{(0.9)(0.01) + \underbrace{(0.1)}(0.99)} = \boxed{\frac{1}{12}}$$

Given low rates of spam, high false positive rate is terrible

General

Risk of a predictor $f(x)$:

$$R(f) = E_{(x,y)\sim D}\left[\ell(f(x),y)\right]$$

Practically: We do not know the distribution D

1. What is the distribution of all images of dogs and cats?
2. What is the distribution over all movie reviews?

Solution: Work with samples from distrib^n D

Empirical

Risk of a predictor $f(x)$ wrt samples S:

$$\hat{R}_S(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i),y_i)$$

where $S = \{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$ are n samples drawn i.i.d. from D

1. $\hat{R}_S(f)$ is an unbiased estimate of $R(f)$

$$\mathbb{E}_D[\hat{R}_S(f)] = \mathbb{E}_D\left[\frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_D[\ell(f(x_i), y_i)]$$

$$= \frac{1}{n}\sum_{i=1}^{n} R(f) = \frac{1}{n} \cdot nR(f)$$

$$= R(f)$$

2. Even $n=1$ is an unbiased estimate, but it has a high variance

$$\mathrm{Var}[\hat{R}_S(f)] = \mathbb{E}\left[\left(\hat{R}_S(f) - R(f)\right)^2\right]$$

$$\propto \frac{1}{n}$$

3. Empirical risk only tells us something about true risk if the estimator f does not depend on the n samples in S

Train / test splits allow:

1. Learn f from training samples to $\min \hat{R}_S(f)$
2. Estimate $R(f)$ via $\hat{R}_S$ on unseen test samples

$$R(f) = \underbrace{\hat{R}_S(f)}_{\text{Empirical Risk}} + \underbrace{\left[R(f) - \hat{R}_S(f)\right]}_{\text{Generalization Gap}}$$

Caveats:

1. Here $\hat{R}_S(f)$ refers to empirical risk on seen training data
2. $R(f)$ is still the true risk and will be approximated by empirical risk on unseen test data in practice

Case: Consider the all-zero house-price predictor

1. $\hat{R}_S(f)$ will be the loss from predicting 0 on training data

$$\hat{R}_S(f) = \frac{1}{n} \sum_{(x_i, y_i) \sim \text{train}} \ell(0, y_i)$$

2. $R(f)$ can be estimated from test data

$$R(f) \approx \hat{R}_{\text{test}}(f) = \frac{1}{n} \sum_{(x_i, y_i) \sim \text{test}} \ell(0, y_i)$$

Assuming train & test are drawn iid from D

1. $\hat{R}_S(f)$ on train will be large

2. $R(f) - \hat{R}_S(f) \approx \hat{R}_{test}(f) - \hat{R}_S(f)$

$$\approx 0$$

Case: The SERIOUS ML predictors

1. $\hat{R}_S(f)$ will be made low via optimization

2. By choosing good model classes, loss functions & optimization procedure we hope $R(f)$ is comparable to $\hat{R}_S(f)$ i.e. $R(f) - \hat{R}_S(f) \approx 0$

- $R(f)$ estimated via test set error will be almost always greater than $\hat{R}_S(f)$ on train

- If $R(f)$ (on unseen data) $< \hat{R}_S(f)$ on train

CAUTION: train / test split may be inappropriate