Discussion 1/19: Probability & Generalization

Q1) 2 red, 3 blue

let A = RV (random variable) color of the ball Alice draws

B = RV color of the ball Bob draws

1) $P(A=\text{red}, B=\text{blue}) =$

* not independent

"given"

$$= \underbrace{P(A=\text{red})}_{\frac{\#\text{ red ball}}{\#\text{ total}} = 2/5} \cdot \underbrace{P(B=\text{blue} \mid A=\text{red})}_{\frac{\#\text{ blue balls}}{\#\text{total} - 1\text{ red Alice drew}} = 3/4}$$

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

2) $P(A=\text{blue} \mid B=\text{blue}) \overset{\text{Bayes rule}}{=} \dfrac{P(B=\text{blue} \mid A=\text{blue})P(A=\text{blue})}{P(B=\text{blue})}$

$$= \frac{P(B=\text{blue} \mid A=\text{blue})P(A=\text{blue})}{P(B=\text{blue} \mid A=\text{blue})P(A=\text{blue}) + P(B=\text{blue} \mid A=\text{red}) \cdot P(A=\text{red})}$$

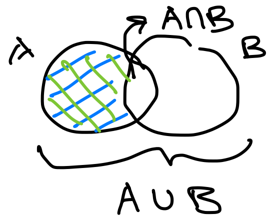
$$= \frac{(2/4)(3/5)}{(2/4)(3/5) + (3/4)(2/5)} = \boxed{\frac{1}{2}}$$

Q2) Events A, B, C

$\cap$: intersection, "and"
$\cup$: union, "or"

a) $P(A) - P(A \cap B) = P(A \cup B) - P(B)$ $\boxed{\text{True}}$

Inclusion-Exclusion Principle



b) $P(A \cup B) \leq P(A) + P(B) - P(A)P(B)$ $\boxed{\text{False}}$

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

$$= P(A) + P(B) - P(A)P(B|A)$$

if $\underline{P(B|A) \geq P(B)}$, then statement b) is true

prove by counterexample : 2 red, 3 blue

$P(B = \text{red}) = 2/5$

$P(B = \text{red} \mid A \text{ red}) = 1/5$

c) $P(A) = P(A \cap C) + P(A \cap \bar{C})$ $\boxed{\text{True}}$

Law of total probability

d) $P(A) = P(A|C) + P(A|\bar{C})$ $\boxed{\text{False}}$

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

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

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

Q3) b) $\frac{P(A|B,C)}{P(A|C)} = \frac{P(B|A,C)}{P(B|C)}$

$$= \frac{\dfrac{P(B,C|A)\cancel{P(A)}}{P(B,C)}}{\dfrac{P(C|A)\cdot\cancel{P(A)}}{P(C)}}$$

$$= \frac{P(B,C|A)}{P(C|A)} \cdot \cancel{\frac{P(C)}{P(B,C)}} \rightarrow \frac{1}{P(B|C)} \text{ b/c } P(B,C) = P(C)\cdot P(B|C)$$

$$= \frac{\cancel{P(C|A)}\cdot P(B|C,A)}{\cancel{P(C|A)}\cdot P(B|C)}$$

$$= \frac{P(B|A,C)}{P(B|C)}$$

Q6) let E: event that email is spam

S: event that email is marked as spam

Given: $P(S|E) = 0.9 \quad P(S|\bar{E}) = 0.1 \quad P(E) = 0.01$

Want: $P(E|S) = ?$

→ Bayes rule

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

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

$$= 0.9 \cdot 0.01 + 0.1 + 0.99$$

$$= 0.108$$

$$P(E|S) = \frac{0.9 \cdot 0.01}{0.108}$$

$$= 0.08\overline{3} = \boxed{\frac{1}{12}}$$

Generalization (Lecture 1):

Given: input space $\mathcal{X}$, output space $\mathcal{Y}$, x, y come from some data distribution D

Want: learn a predictor $f(x): \mathcal{X} \to \mathcal{Y}$ that minimizes some loss function $\ell(f(x), y)$

Ex: square footage → house price
health history → risk of disease
photo → is this a bicycle?

Risk: $R(f) = \mathbb{E}_{(x,y)\sim D}[\ell(f(x), y)]$
↳ don't have infinite data

Empirical Risk: estimate $R(f)$ with sample average for data points $S = \{(x_1, y_1) \dots (x_n, y_n)\} \overset{iid}{\sim} D$

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

- $\hat{R}_S(f)$ is unbiased estimator of $R(f)$

$$\mathbb{E}_D[\hat{R}_S(f)] = R(f)$$

Q: How close is $\hat{R}_S(f)$ to $R(f)$?

Given training data S, to find best predictor, choose f that minimizes $\hat{R}_S(f)$.

- We want f to generalize beyond S to all similar types of data.
  ↳ from same data distribution D

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

How can we estimate $R(f)$?

Test error: Separate test dataset S', NOT training set, but drawn from same distribution.

$$R(f) \approx \frac{1}{m} \sum_{i=1}^{m} \ell(f(x_i'), y_i')$$
$$= \hat{R}_{S'}(f)$$

Q: Why can't we use S to estimate $R(f)$.

$\hat{R}_S(f)$ is only unbiased if f does not depend on S.

Ex: f: if x is in S, output correct y; otherwise, output nonsense

$\hat{R}_S(f) = 0$, $R(f)$ high

Q: Why don't we train f using both S and S'.

- Learn f from $\hat{R}_S(f)$
- Estimate $R(f)$ using $\hat{R}_{S'}(f)$