Plan du cours

1. Introduction générale
2. Preuves (revue)
3. Algèbre linéaire (revue)
4. Optimisation (revue)
5. Optimisation sous contraintes
6. Probabilités (revue)
7. Statistique
8. **Théorie de l'apprentissage**
9. **Optimisation pour l'apprentissage**

Crédits

Transparents repris et modifié de

- Taide Ding, Fereshte Khani, Stanford CS229 probability theory review, April 2020
- Wikipedia

Elements of Probability

Sample Space Ω

$\{HH, HT, TH, TT\}$

Event $A \subseteq \Omega$

$\{HH, HT\}$, Ω

Event Space $\mathcal{F}$

Probability Measure $P : \mathcal{F} \to \mathbb{R}$

$P(A) \geq 0 \quad \forall A \in \mathcal{F}$

$P(\Omega) = 1$

countable

If $A_1, A_2, \ldots$ disjoint set of events ($A_i \cap A_j = \emptyset$ when $i \neq j$), then

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

Conditional Probability and Bayes' Rule

For any events A, B such that $P(B) \neq 0$, we define:

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}$$

Let's apply conditional probability to obtain **Bayes' Rule**!

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$
$$= \boxed{\frac{P(B)P(A \mid B)}{P(A)}}$$

Conditioned Bayes' Rule: given events A, B, C,

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

See Appendix for proof :)

Law of Total Probability

Let $B_1, \ldots, B_n$ be n disjoint events whose union is the entire sample space. Then, for any event A,

$$\begin{aligned} P(A) &= \sum_{i=1}^{n} P(A \cap B_i) \\ &= \sum_{i=1}^{n} P(A \mid B_i) P(B_i) \end{aligned}$$

We can then write Bayes' Rule as:

$$\begin{aligned} P(B_k \mid A) &= \frac{P(B_k) P(A \mid B_k)}{P(A)} \\ &= \boxed{\frac{P(B_k) P(A \mid B_k)}{\sum_{i=1}^{n} P(A \mid B_i) P(B_i)}} \end{aligned}$$

Example

Treasure chest **A** holds 100 gold coins. Treasure chest **B** holds 60 gold and 40 silver coins.
Choose a treasure chest uniformly at random, and pick a coin from that chest uniformly at random. If the coin is gold, then what is the probability that you chose chest **A**? [1]

[1] Question based on slides by Koochak & Irvin

Example

Treasure chest **A** holds 100 gold coins. Treasure chest **B** holds 60 gold and 40 silver coins.
Choose a treasure chest uniformly at random, and pick a coin from that chest uniformly at random. If the coin is gold, then what is the probability that you chose chest **A**? [1]
Solution:

$$\begin{aligned} P(A \mid G) &= \frac{P(A)P(G \mid A)}{P(A)P(G \mid A) + P(B)P(G \mid B)} \\ &= \frac{0.5 \times 1}{0.5 \times 1 + 0.5 \times 0.6} \\ &= \boxed{0.625} \end{aligned}$$

[1] Question based on slides by Koochak & Irvin

Chain Rule

For any n events $A_1, ..., A_n$, the joint probability can be expressed as a product of conditionals:

$$P(A_1 \cap A_2 \cap ... \cap A_n)$$
$$= P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_2 \cap A_1)...P(A_n \mid A_{n-1} \cap A_{n-2} \cap ... \cap A_1)$$

Independence

Events A, B are independent if

$$P(AB) = P(A)P(B)$$

We denote this as $A \perp B$. From this, we know that if $A \perp B$,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Implication: If two events are independent, observing one event does not change the probability that the other event occurs.
In general: events $A_1, ..., A_n$ are **mutually independent** if

$$P(\bigcap_{i \in S} A_i) = \prod_{i \in S} P(A_i)$$

for any subset $S \subseteq \{1, ..., n\}$.

Random Variables

real

- A **random variable** X maps outcomes to real values.
- X takes on values in $Val(X) \subseteq \mathbb{R}$.
- $X = k$ is the event that random variable X takes on value k.

Discrete RVs:

- $Val(X)$ is a set
- $P(X = k)$ can be nonzero

Continuous RVs:

- $Val(X)$ is a range
- $P(X = k) = 0$ for all k. $P(a \leq X \leq b)$ can be nonzero.

Mixed RVs

Probability Mass Function (PMF)

Given a **discrete** RV X, a PMF maps values of X to probabilities.

$$p_X(x) := P(X = x)$$

For a valid PMF, $\sum_{x \in Val(x)} p_X(x) = 1$.

Cumulative Distribution Function (CDF)

A CDF maps a continuous RV to a function $\mathbb{R} \to [0, 1]$

$$F_X(x) := P(X \leq x)$$

A CDF must fulfill the following:

- $\lim_{x \to -\infty} F_X(x) = 0$
- $\lim_{x \to \infty} F_X(x) = 1$
- If $a \leq b$, then $F_X(a) \leq F_X(b)$ (i.e. CDF must be nondecreasing)

Also note: $P(a \leq X \leq b) = F_X(b) - F_X(a)$.

Probability Density Function (PDF)

PDF of a continuous RV is simply the derivative of the CDF.

$$f_X(x) := \frac{dF_X(x)}{dx}$$

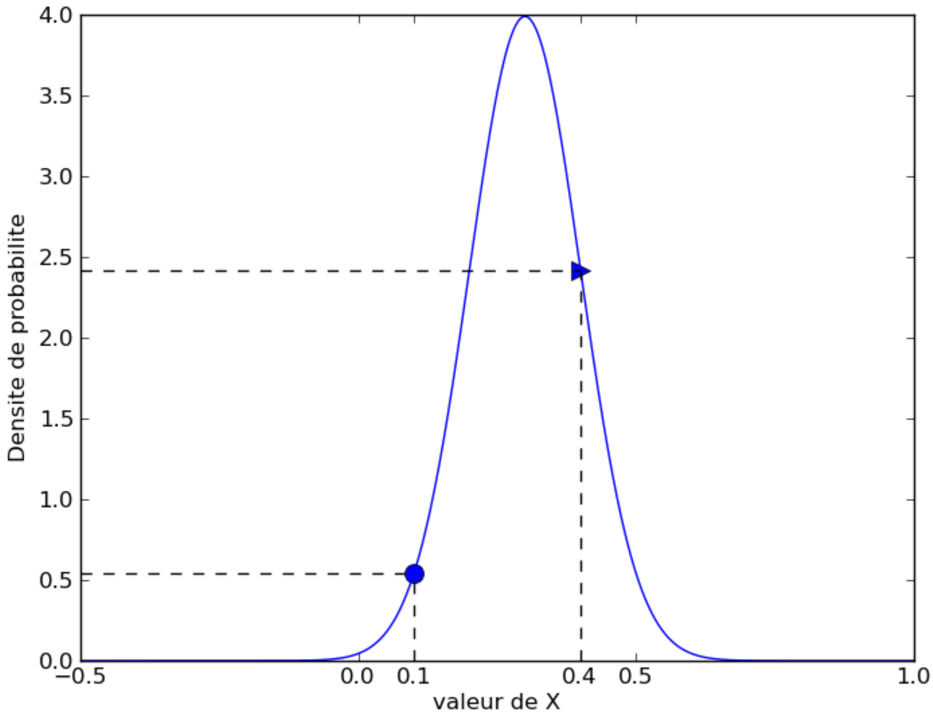
Thus,

$$P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx$$

A valid PDF must be such that

- for all real numbers x, $f_X(x) \geq 0$.
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$

On a représente sur le graphe ci-dessous la densité de probabilité d'une variable aléatoire réelle X.



1) La densité de probabilité de X prend une valeur supérieure 1 en $X = 0.4$. Cela vous parait-il normal ? Justifiez votre réponse.

Soit x une réalisation de X.

2) Quelle est la probabilité d'avoir $x = 0.1$? Quelle est la probabilité d'avoir $x = 0.4$? Est il plus probable d'observer $x = 0.4$ ou $x = 0.1$? A quel point (approximativement) ?

Expectation

Let g be an arbitrary real-valued function.

- If X is a discrete RV with PMF p_X:

$$\mathbb{E}[g(X)] := \sum_{x \in Val(X)} g(x) p_X(x)$$

- If X is a continuous RV with PDF f_X:

$$\mathbb{E}[g(X)] := \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Intuitively, expectation is a weighted average of the values of $g(x)$, weighted by the probability of x.

Properties of Expectation

For any constant $a \in \mathbb{R}$ and arbitrary real function f:

- $\mathbb{E}[a] = a$
- $\mathbb{E}[af(X)] = a\mathbb{E}[f(X)]$

Linearity of Expectation
Given n real-valued functions $f_1(X), ..., f_n(X)$,

$$\mathbb{E}[\sum_{i=1}^{n} f_i(X)] = \sum_{i=1}^{n} \mathbb{E}[f_i(X)]$$

Law of Total Expectation
Given two RVs X, Y:

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$$

N.B. $\mathbb{E}[X \mid Y] = \sum_{x \in Val(x)} x p_{X|Y}(x|y)$ is a function of Y.
See Appendix for details :)

Example of Law of Total Expectation

El Goog sources two batteries, A and B, for its phone. A phone with battery A runs on average 12 hours on a single charge, but only 8 hours on average with battery B. El Goog puts battery A in 80% of its phones and battery B in the rest. If you buy a phone from El Goog, how many hours do you expect it to run on a single charge?

Example of Law of Total Expectation

El Goog sources two batteries, A and B, for its phone. A phone with battery A runs on average 12 hours on a single charge, but only 8 hours on average with battery B. El Goog puts battery A in 80% of its phones and battery B in the rest. If you buy a phone from El Goog, how many hours do you expect it to run on a single charge?

Solution: Let L be the time your phone runs on a single charge. We know the following:

- $p_X(A) = 0.8$, $p_X(B) = 0.2$,
- $\mathbb{E}[L \mid A] = 12$, $\mathbb{E}[L \mid B] = 8$.

Then, by Law of Total Expectation,

$$\begin{aligned}
\mathbb{E}[L] = \mathbb{E}[\mathbb{E}[L \mid X]] &= \sum_{X \in \{A,B\}} \mathbb{E}[L \mid X] p_X(X) \\
&= \mathbb{E}[L \mid A] p_X(A) + \mathbb{E}[L \mid B] p_X(B) \\
&= 12 \times 0.8 + 8 \times 0.2 = \boxed{11.2}
\end{aligned}$$

Variance

The **variance** of a RV X measures how concentrated the distribution of X is around its mean.

$$\begin{aligned} Var(X) &:= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \end{aligned}$$

Interpretation: $Var(X)$ is the expected deviation of X from $\mathbb{E}[X]$.
Properties: For any constant $a \in \mathbb{R}$, real-valued function $f(X)$

- $Var[a] = 0$
- $Var[af(X)] = a^2 Var[f(X)]$

Example Distributions

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, ..., n$	np	$np(1-p)$
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, ...$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$\frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, ...$	λ	λ
$Uniform(a, b)$	$\frac{1}{b-a}$ for all $x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for all $x \in (-\infty, \infty)$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ for all $x \geq 0, \lambda \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Joint and Marginal Distributions

- **Joint PMF** for discrete RV's X, Y:

$$p_{XY}(x, y) = P(X = x, Y = y)$$

  Note that $\sum_{x \in Val(X)} \sum_{y \in Val(Y)} p_{XY}(x, y) = 1$
- **Marginal PMF** of X, given joint PMF of X, Y:

$$p_X(x) = \sum_y p_{XY}(x, y)$$

- **Joint PDF** for continuous X, Y:

$$f_{XY}(x, y) = \frac{\delta^2 F_{XY}(x, y)}{\delta x \delta y}$$

  Note that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1$
- **Marginal PDF** of X, given joint PDF of X, Y:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$

Joint and Marginal Distributions for Multiple RVs

- **Joint PMF** for discrete RV's $X_1, ..., X_n$:

$$p(x_1, ..., x_n) = P(X_1 = x_1, ..., X_n = x_n)$$

  Note that $\sum_{x_1} \sum_{x_2} ... \sum_{x_n} p(x_1, ..., x_n) = 1$
- **Marginal PMF** of X_1, given joint PMF of $X_1, ..., X_n$:

$$p_{X_1}(x_1) = \sum_{x_2} ... \sum_{x_n} p(x_1, ..., x_n)$$

- **Joint PDF** for continuous RV's $X_1, ..., X_n$:

$$f(x_1, ..., x_n) = \frac{\delta^n F(x_1, ...x_n)}{\delta x_1 \delta x_2 ... \delta x_n}$$

  Note that $\int_{x_1} \int_{x_2} ... \int_{x_n} f(x_1, ..., x_n) dx_1 ... dx_n = 1$
- **Marginal PDF** of X_1, given joint PDF of $X_1, ..., X_n$:

$$f_{X_1}(x_1) = \int_{x_2} ... \int_{x_n} f(x_1, ..., x_n) dx_2 ... dx_n$$

Expectation for multiple random variables

Given two RV's X, Y and a function $g : \mathbb{R}^2 \to \mathbb{R}$ of X, Y,

- for discrete X, Y:

$$\mathbb{E}[g(X, Y)] := \sum_{x \in Val(x)} \sum_{y \in Val(y)} g(x, y) p_{XY}(x, y)$$

- for continuous X, Y:

$$\mathbb{E}[g(X, Y)] := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dxdy$$

These definitions can be extended to multiple random variables in the same way as in the previous slide. For example, for n continuous RV's $X_1, .., X_n$ and function $g : \mathbb{R}^n \to \mathbb{R}$:

$$\mathbb{E}[g(X)] = \int \int ... \int g(x_1, ..., x_n) f_{X_1,...,X_n}(x_1, ..., x_n) dx_1, ..., dx_n$$

Covariance

Intuitively: measures how much one RV's value tends to move with another RV's value. For RV's X, Y:

$$
\begin{aligned}
Cov[X, Y] &:= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}
$$

- If $Cov[X, Y] < 0$, then X and Y are negatively correlated
- If $Cov[X, Y] > 0$, then X and Y are positively correlated
- If $Cov[X, Y] = 0$, then X and Y are uncorrelated

Properties Involving Covariance

- If $X \perp Y$, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Thus,

$$Cov[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$$

  This is unidirectional! $Cov[X, Y] = 0$ **does not imply** $X \perp Y$
- **Variance of two variables:**

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$$

  i.e. if $X \perp Y$, $Var[X + Y] = Var[X] + Var[Y]$.
- **Special Case:**

$$Cov[X, X] = \mathbb{E}[XX] - \mathbb{E}[X]\mathbb{E}[X] = Var[X]$$

Variance of a sum

$$\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{V}(X_i) + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j)$$

Exercice : espérance et variance de la moyenne de n variables aléatoires i.id. ?

Law of total variance

$$\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathrm{E}[Y \mid X]).$$

Exercice : On a une procédure aléatoire pour entrainer un classificateur binaire, dont on obtient n échantillons (n classifieurs). Pour tester la qualité de la procédure d'entrainement, on a une procédure aléatoire de test qui produit une erreur de classification et qu'on applique m fois sur chacun des n classificateurs entrainé . Comment mesurer la variance de l'erreur de classification (par exemple pour savoir si elle est significativement en dessous du hasard) à partir des erreurs de classifications $(e_{i,j})_{1 \le i \le n, 1 \le j \le m}$?

Conditional distributions for RVs

Works the same way with RV's as with events:

- For discrete X, Y:

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)}$$

- For continuous X, Y:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

- In general, for continuous $X_1, ..., X_n$:

$$f_{X_1|X_2,...,X_n}(x_1|x_2, ..., x_n) = \frac{f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)}{f_{X_2,...,X_n}(x_2, ..., x_n)}$$

Bayes' Rule for RVs

Also works the same way for RV's as with events:

- For discrete X, Y:

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)p_Y(y)}{\sum_{y' \in Val(Y)} p_{X|Y}(x|y')p_Y(y')}$$

- For continuous X, Y:

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y')f_Y(y')dy'}$$

Chain Rule for RVs

Also works the same way as with events:

$$\begin{aligned} f(x_1, x_2, ..., x_n) &= f(x_1) f(x_2 | x_1) ... f(x_n | x_1, x_2, ..., x_{n-1}) \\ &= f(x_1) \prod_{i=2}^{n} f(x_i | x_1, ..., x_{i-1}) \end{aligned}$$

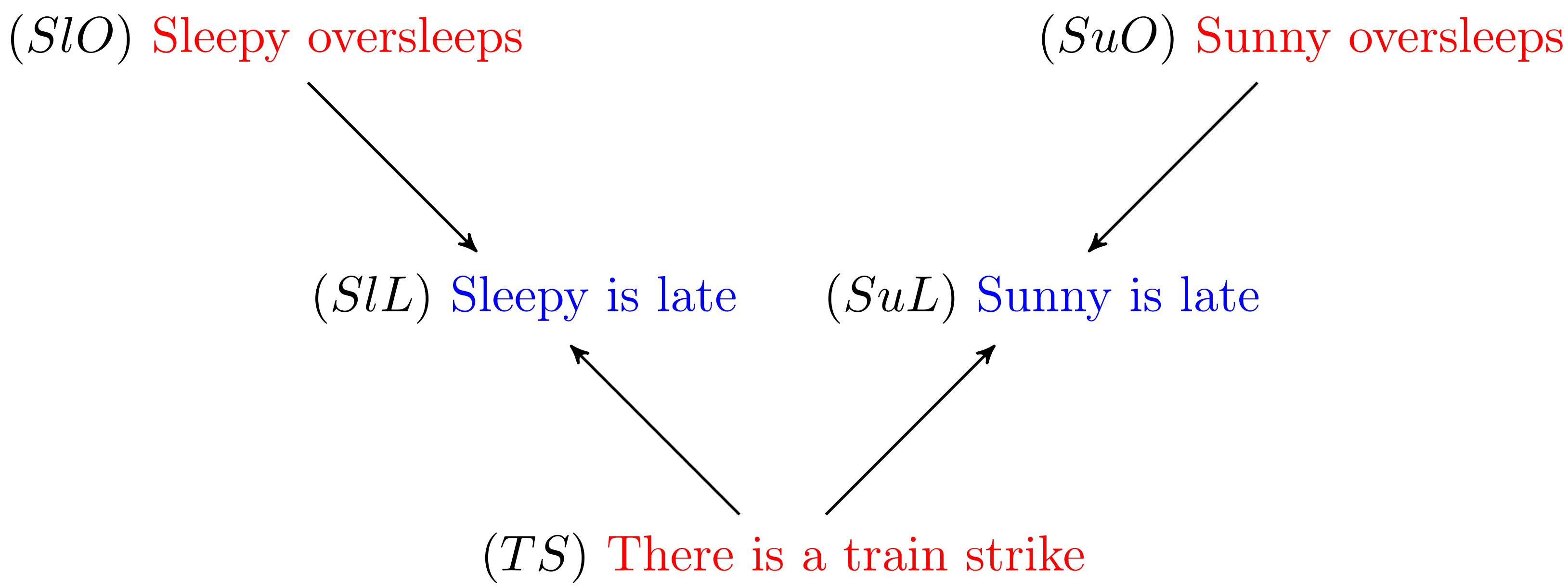
Independence for RVs

- For $X \perp Y$ to hold, it must be that $F_{XY}(x, y) = F_X(x)F_Y(y)$ **FOR ALL VALUES** of x, y.

- Since $f_{Y|X}(y|x) = f_Y(y)$ if $X \perp Y$, chain rule for mutually independent $X_1, ..., X_n$ is:

$$f(x_1, ..., x_n) = f(x_1)f(x_2)...f(x_n) = \prod_{i=1}^{n} f(x_i)$$

(very important assumption for a Naive Bayes classifier!)



SlO, SuO, SlL, SuL and TS binary random variables

$$P(SlL = 1 | SlO = a, TS = b) = a \vee b,$$

$$P(SuL = 1 | TS = b, SuO = c) = b \vee c,$$

$$l = P(SlO = 1),\ u = P(SuO = 1) \text{ and } t = P(TS = 1).$$

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | X_{\pi_{i,1}}, X_{\pi_{i,2}}, ..., X_{\pi_{i,n_i}}),$$

where $\{\pi_{i,1}, \pi_{i,2}, ..., \pi_{i,n_i}\}$ is the set of the parents of node i in the graph.

This formula is often abridged into :

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i | X_{\pi_i}).$$

1. Express the factorization of $P(SlL, SuL, SlO, SuO, TS)$ in the graphical model
2. Is the joint probability entirely specified if we know the values of l, u and t?
3. Compute $P(TS = 1 | SlL = 1)$ as a function of l, u and t.
4. Compute $P(SlO = 1 | SlL = 1)$ as a function of l, u and t.
5. Compute $P(TS = 1 | SlL = 1, SuL = 1)$ as a function of l, u and t.
6. Compute $P(SlO = 1 | SlL = 1, SuL = 1)$ as a function of l, u and t.
7. Suppose now that $l = 0.5$, $t = 0.1$ and that we observe that *Sleepy is late*. What is the most probable : that *there is a train strike* or that *Sleepy overslept*?
8. Same question when we suppose in addition that $u = 0.01$ and that we observe that *Sunny is late* too.
9. What happens if we take $l = 0.5$, $t = 0.1$ and $u = 0.2$?

Appendix: More on Total Expectation

Why is $\mathbb{E}[X|Y]$ a function of Y? Consider the following:

- $\mathbb{E}[X|Y = y]$ is a scalar that only depends on y.
- Thus, $\mathbb{E}[X|Y]$ is a random variable that only depends on Y. Specifically, $\mathbb{E}[X|Y]$ is a function of Y mapping $Val(Y)$ to the real numbers.

An example: Consider RV X such that

$$X = Y^2 + \epsilon$$

such that $\epsilon \sim \mathcal{N}(0, 1)$ is a standard Gaussian. Then,

- $\mathbb{E}[X|Y] = Y^2$
- $\mathbb{E}[X|Y = y] = y^2$

Appendix: More on Total Expectation

A derivation of Law of Total Expectation for discrete X, Y:[3]

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[\sum_x xP(X = x \mid Y)] \tag{1}$$

$$= \sum_y \sum_x xP(X = x \mid Y)P(Y = y) \tag{2}$$

$$= \sum_y \sum_x xP(X = x, Y = y) \tag{3}$$

$$= \sum_x x \sum_y P(X = x, Y = y) \tag{4}$$

$$= \sum_x xP(X = x) = \boxed{\mathbb{E}[X]} \tag{5}$$

where (1), (2), and (5) result from the definition of expectation, (3) results from the definition of cond. prob., and (5) results from marginalizing out Y.

[3]from slides by Koochak & Irvin

Appendix: A proof of Conditioned Bayes Rule

Repeatedly applying the definition of conditional probability, we have: [4]

$$\begin{aligned}
\frac{P(b|a,c)P(a|c)}{P(b|c)} &= \frac{P(b,a,c)}{P(a,c)} \cdot \frac{P(a|c)}{P(b|c)} \\
&= \frac{P(b,a,c)}{P(a,c)} \cdot \frac{P(a,c)}{P(b|c)P(c)} \\
&= \frac{P(b,a,c)}{P(b|c)P(c)} \\
&= \frac{P(b,a,c)}{P(b,c)} \\
&= P(a|b,c)
\end{aligned}$$

[4] from slides by Koochak & Irvin