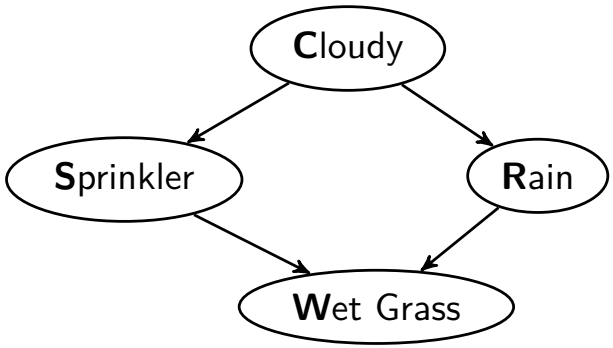
Inférence sur les programmes probabilistes

Dérivation d'un algorithme d'inférence variationnelle

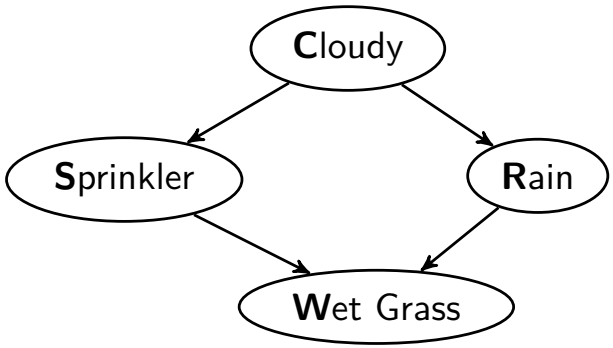
Raphaël Monat

Stage effectué à l'université d'Oxford, Royaume-Uni
Sous la direction de Hongseok Yang

1. Modèles probabilistes
2. Programmes probabilistes
3. Contribution
4. Conclusion

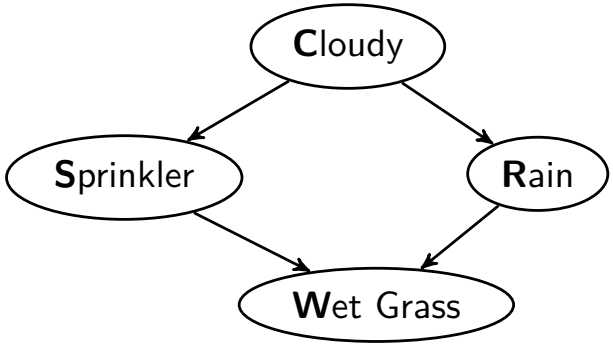

Cloudy
Sprinkler
Rain
Wet Grass

$\Pr(C = T)$	$\Pr(C = F)$
$0,4$	$0,6$



$\Pr(C = T)$	$\Pr(C = F)$
$0,4$	$0,6$

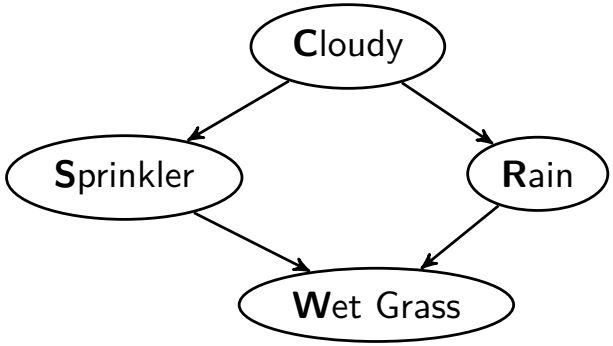
C	$\Pr(S = T)$	$\Pr(S = F)$
T	$0,1$	$0,9$
F	$0,5$	$0,5$



C	$\Pr(R = T)$	$\Pr(R = F)$
T	$0,8$	$0,2$
F	$0,2$	$0,8$

$\Pr(C = T)$	$\Pr(C = F)$
$0,4$	$0,6$

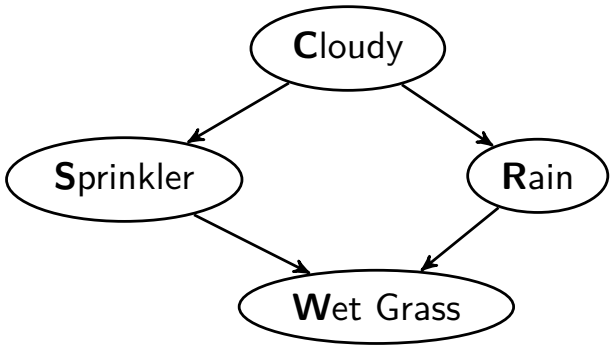
C	$\Pr(S = T)$	$\Pr(S = F)$
T	$0,1$	$0,9$
F	$0,5$	$0,5$



C	$\Pr(R = T)$	$\Pr(R = F)$
T	$0,8$	$0,2$
F	$0,2$	$0,8$

$\Pr(C = T)$	$\Pr(C = F)$
$0,4$	$0,6$

C	$\Pr(S = T)$	$\Pr(S = F)$
T	$0,1$	$0,9$
F	$0,5$	$0,5$

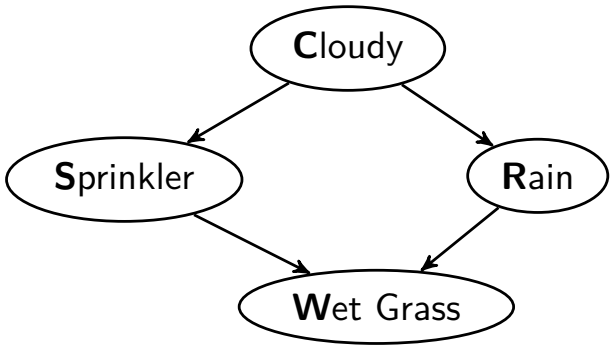


C	$\Pr(R = T)$	$\Pr(R = F)$
T	$0,8$	$0,2$
F	$0,2$	$0,8$

S	R	$\Pr(W = T)$	$\Pr(W = F)$
T	T	$0,99$	$0,01$
T	F	$0,9$	$0,1$
F	T	$0,9$	$0,1$
F	F	0	1

$\Pr(C = T)$	$\Pr(C = F)$
$0,4$	$0,6$

C	$\Pr(S = T)$	$\Pr(S = F)$
T	$0,1$	$0,9$
F	$0,5$	$0,5$



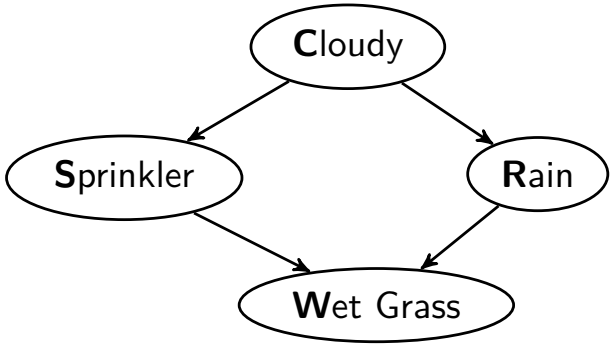
C	$\Pr(R = T)$	$\Pr(R = F)$
T	$0,8$	$0,2$
F	$0,2$	$0,8$

S	R	$\Pr(W = T)$	$\Pr(W = F)$
T	T	$0,99$	$0,01$
T	F	$0,9$	$0,1$
F	T	$0,9$	$0,1$
F	F	0	1

$$\Pr(W, S, R, C) = \Pr(W|S, R) \Pr(S|C) \Pr(R|C) \Pr(C)$$

$\Pr(C = T)$	$\Pr(C = F)$
$0,4$	$0,6$

C	$\Pr(S = T)$	$\Pr(S = F)$
T	$0,1$	$0,9$
F	$0,5$	$0,5$



C	$\Pr(R = T)$	$\Pr(R = F)$
T	$0,8$	$0,2$
F	$0,2$	$0,8$

S	R	$\Pr(W = T)$	$\Pr(W = F)$
T	T	$0,99$	$0,01$
T	F	$0,9$	$0,1$
F	T	$0,9$	$0,1$
F	F	0	1

$$\Pr(W = w, S = s, R = r, C = c)$$
$$= \Pr(W = w|S = s, R = r) \Pr(S = s|C = c) \Pr(R = r|C = c) \Pr(C = c)$$

Faire des requêtes auprès du modèle.

$$
\begin{aligned}
&\Pr(R = T | W = T, S = T) \\
&= \frac{\Pr(R = T, W = T, S = T)}{\Pr(W = T, S = T)} \\
&= \frac{\sum_{c \in \{T,F\}} \Pr(W = T, S = T, R = T, C = c)}{\sum_{r,c \in \{T,F\}^2} \Pr(W = T, S = T, R = r, C = c)} \\
&= \frac{0,03168 + 0,0594}{0,03168 + 0,0594 + 0,0072 + 0,216} \simeq 28,98\%
\end{aligned}
$$

Formulation générale du problème de l'inférence :

$$\Pr(\theta|D,\mathcal{H}) = \frac{\Pr(D|\theta,\mathcal{H})\Pr(\theta|\mathcal{H})}{\Pr(D|\mathcal{H})}$$

- $\Pr(\theta|D,\mathcal{H})$ probabilité *a posteriori* de θ sachant les données D et le modèle $\mathcal{H}$;

Formulation générale du problème de l'inférence :

$$\Pr(\theta|D,\mathcal{H}) = \frac{\Pr(D|\theta,\mathcal{H})\Pr(\theta|\mathcal{H})}{\Pr(D|\mathcal{H})}$$

- $\Pr(\theta|D,\mathcal{H})$ probabilité *a posteriori* de θ sachant les données D et le modèle $\mathcal{H}$;
- $\Pr(D|\theta,\mathcal{H})$ fonction de vraisemblance de θ dans le modèle $\mathcal{H}$;

Formulation générale du problème de l'inférence :

$$\Pr(\theta|D,\mathcal{H}) = \frac{\Pr(D|\theta,\mathcal{H})\Pr(\theta|\mathcal{H})}{\Pr(D|\mathcal{H})}$$

- $\Pr(\theta|D,\mathcal{H})$ probabilité *a posteriori* de θ sachant les données D et le modèle $\mathcal{H}$;
- $\Pr(D|\theta,\mathcal{H})$ fonction de vraisemblance de θ dans le modèle $\mathcal{H}$;
- $\Pr(\theta|\mathcal{H})$ probabilité *a priori* de θ ;

Formulation générale du problème de l'inférence :

$$\Pr(\theta|D,\mathcal{H}) = \frac{\Pr(D|\theta,\mathcal{H})\Pr(\theta|\mathcal{H})}{\Pr(D|\mathcal{H})}$$

- $\Pr(\theta|D,\mathcal{H})$ probabilité *a posteriori* de θ sachant les données D et le modèle $\mathcal{H}$;
- $\Pr(D|\theta,\mathcal{H})$ fonction de vraisemblance de θ dans le modèle $\mathcal{H}$;
- $\Pr(\theta|\mathcal{H})$ probabilité *a priori* de θ ;
- $\Pr(D|\mathcal{H})$ probabilité marginale de D.

Méthode exacte : retour à la probabilité jointe

$$\Pr(R = T | W = T, S = T) = \frac{\sum_{c \in \{T,F\}} \Pr(W = T, S = T, R = T, C = c)}{\sum_{r,c \in \{T,F\}^2} \Pr(W = T, S = T, R = r, C = c)}$$

Remarque (Complexité)

Le problème de l'inférence est NP-Complet dans le cas des réseaux bayésiens.

Inférence variationnelle : approximer $p(\theta|D)$ par $q(\theta) \in \mathcal{Q}$

Inférence variationnelle : approximer $p(\theta|D)$ par $q(\theta) \in \mathcal{Q}$

Définition (Kullback-Leibler divergence)

p et q deux probabilités définies sur le même espace.

$$\mathrm{KL}(p||q) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$$

Inférence variationnelle : approximer $p(\theta|D)$ par $q(\theta) \in \mathcal{Q}$

Définition (Kullback-Leibler divergence)

p et q deux probabilités définies sur le même espace.

$$\mathsf{KL}(p||q) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$$

Objectif

$$\operatorname*{argmin}_{q \in \mathcal{Q}} \mathsf{KL}(q(\theta)||p(\theta|D))$$

Inférence variationnelle : approximer $p(\theta|D)$ par $q(\theta) \in \mathcal{Q}$

Définition (Kullback-Leibler divergence)

p et q deux probabilités définies sur le même espace.

$$\mathsf{KL}(p||q) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$$

Objectif

$$\operatorname*{argmin}_{q \in \mathcal{Q}} \mathsf{KL}(q(\theta)||p(\theta|D))$$

Problème

Calcul de $p(\theta|D)$

Définition ("Evidence Lower Bound")

$$\mathcal{L}_{vi}(q) := \log p(D) - \mathsf{KL}(q(\theta)||p(\theta|D))$$

$$\begin{aligned}
\mathcal{L}_{vi}(q) &= \log p(D) \cdot \int q(\theta)\,\mathrm{d}\theta - \int q(\theta) \log \frac{q(\theta)}{p(\theta|D)}\,\mathrm{d}\theta \\
&= \int q(\theta) \log p(D)\,\mathrm{d}\theta + \int q(\theta) \log \frac{p(\theta|D)}{q(\theta)}\,\mathrm{d}\theta \\
&= \int q(\theta) \log \frac{p(\theta, D)}{q(\theta)}\,\mathrm{d}\theta
\end{aligned}$$

$$\operatorname*{argmin}_{q \in \mathcal{Q}} \mathsf{KL}(q||p) = \operatorname*{argmax}_{q \in \mathcal{Q}} \mathcal{L}_{vi}(q)$$

But : simplifier le développement de nouveaux modèles

- Modèles $\rightarrow$ programmes ;
- Plus expressif que les modèles probabilistes ;
- Algorithme d'inférence général (pas d'intervention de l'utilisateur) ;

Anglican = Clojure +

- `(defquery myquery [args] <body>)` ;

Anglican = Clojure +

- `(defquery myquery [args] <body>);`
- `(sample <dist>);`

Anglican = Clojure +

- `(defquery myquery [args] <body>);`
- `(sample <dist>);`
- `(observe <dist> <value>);`

Anglican = Clojure +

- `(defquery myquery [args] <body>);`
- `(sample <dist>);`
- `(observe <dist> <value>);`
- `(predict <expr>).`

```
(defquery bayes-net [sprinkler wet-grass]
 (let [is-cloudy (sample (flip 0.4))
       is-raining (cond (= is-cloudy true )
                        (sample (flip 0.8))
                        (= is-cloudy false)
                        (sample (flip 0.2)))
       sprinkler-dist (cond (= is-cloudy true)
                            (flip 0.1)
                            (= is-cloudy false)
                            (flip 0.5))
       wet-grass-dist (cond
                        (and (= sprinkler true)
                             (= is-raining true))
                        (flip 0.99)
                        (and (= sprinkler false)
                             (= is-raining false))
                        (false)
                        (or  (= sprinkler true)
                             (= is-raining true))
                        (flip 0.9))]
   (observe sprinkler-dist sprinkler)
   (observe wet-grass-dist wet-grass)
   (predict :is-raining is-raining)))
```

Définition

Une *densité de transition* sur $(S, \mathcal{F}, \mu)$ est une fonction mesurable $k : S \times S \to \mathbb{R}_+$ telle que :

1. $\forall s \in S$,
$$\int_S k(s, s')\, \mu(\mathrm{d}s') = 1;$$
2. pour tout $S_0 \subseteq S$ mesurable, et pour tout $r \in \mathbb{R}_+$, l'ensemble suivant est mesurable :
$$\left\{ s \;\middle|\; \int_{S_0} k(s, s')\, \mu(\mathrm{d}s') < r \right\}$$

Définition

Un système de transition probabiliste (abrégé en PTS) est un 7-uplet $M = (S, \mathcal{F}, \mu, f, \delta, \psi, F)$ avec :

- (États) $(S, \mathcal{F}, \mu)$ est un espace mesuré tel que $\mu(S) < +\infty$;

Définition

Un système de transition probabiliste (abrégé en PTS) est un 7-uplet $M = (S, \mathcal{F}, \mu, f, \delta, \psi, F)$ avec :

- (États) $(S, \mathcal{F}, \mu)$ est un espace mesuré tel que $\mu(S) < +\infty$;
- (États initiaux) $f : S \to \mathbb{R}_+$ est une fonction mesurable avec

$$\int_S f(s)\,\mu(\mathrm{d}s) = 1;$$

Définition

Un système de transition probabiliste (abrégé en PTS) est un 7-uplet $M = (S, \mathcal{F}, \mu, f, \delta, \psi, F)$ avec :

- (États) $(S, \mathcal{F}, \mu)$ est un espace mesuré tel que $\mu(S) < +\infty$;
- (États initiaux) $f : S \to \mathbb{R}_+$ est une fonction mesurable avec

$$\int_S f(s)\,\mu(\mathrm{d}s) = 1;$$

- (Transition) δ est une densité de transition sur $(S, \mathcal{F}, \mu)$;

Définition

Un système de transition probabiliste (abrégé en PTS) est un 7-uplet $M = (S, \mathcal{F}, \mu, f, \delta, \psi, F)$ avec :

- (États) $(S, \mathcal{F}, \mu)$ est un espace mesuré tel que $\mu(S) < +\infty$;
- (États initiaux) $f : S \to \mathbb{R}_+$ est une fonction mesurable avec

$$\int_S f(s)\,\mu(\mathrm{d}s) = 1;$$

- (Transition) δ est une densité de transition sur $(S, \mathcal{F}, \mu)$;
- (Score) $\psi : S \to \mathbb{R}_+$ est une fonction mesurable ;

Définition

Un système de transition probabiliste (abrégé en PTS) est un 7-uplet $M = (S, \mathcal{F}, \mu, f, \delta, \psi, F)$ avec :

- (États) $(S, \mathcal{F}, \mu)$ est un espace mesuré tel que $\mu(S) < +\infty$;
- (États initiaux) $f : S \to \mathbb{R}_+$ est une fonction mesurable avec

$$\int_S f(s)\,\mu(\mathrm{d}s) = 1;$$

- (Transition) δ est une densité de transition sur $(S, \mathcal{F}, \mu)$;
- (Score) $\psi : S \to \mathbb{R}_+$ est une fonction mesurable ;
- (États finaux) $F \subseteq S$ est un sous-ensemble mesurable de S.

```
(defquery coin []
  (let [l1 is-fair (sample (flip 0.9))
        l2 coin (if is-fair
                    (flip 0.5)
                    (flip 0.95))]
    l3 (observe coin 1)
    l4 (observe coin 1)
    l5 (predict is-fair))l6)
```

Exemple

$$S = [\![1, 6]\!] \times \{0, 1\}$$

```
(defquery coin []
  (let [l1 is-fair
        (sample (flip 0.9))
        l2 coin
        (if is-fair (flip 0.5)
            (flip 0.95))]
    l3 (observe coin 1)
    l4 (observe coin 1)
    l5 (predict is-fair)) l6)
```

Exemple

```
(defquery coin []
  (let [l1 is-fair
        (sample (flip 0.9))
        l2 coin
        (if is-fair (flip 0.5)
            (flip 0.95))]
    l3 (observe coin 1)
    l4 (observe coin 1)
    l5 (predict is-fair)) l6 )
```

$S = [\![1, 6]\!] \times \{0, 1\}$
$f((1, 0)) := 1$

Exemple

```
(defquery coin []
  (let [l1 is-fair
        (sample (flip 0.9))
        l2 coin
        (if is-fair (flip 0.5)
            (flip 0.95))]
    l3 (observe coin 1)
    l4 (observe coin 1)
    l5 (predict is-fair)) l6 )
```

$$S = [\![1, 6]\!] \times \{0, 1\}$$
$$f((1, 0)) := 1$$
$$F = \{(6, 0), (6, 1)\}$$

Exemple

```
(defquery coin []
  (let [l1 is-fair
        (sample (flip 0.9))
        l2 coin
        (if is-fair (flip 0.5)
            (flip 0.95))]
    l3 (observe coin 1)
    l4 (observe coin 1)
    l5 (predict is-fair)) l6 )
```

$$S = [\![1, 6]\!] \times \{0, 1\}$$
$$f((1, 0)) := 1$$
$$F = \{(6, 0), (6, 1)\}$$

$$\delta((1, 0), (2, 0)) = 0, 1$$

Exemple

```
(defquery coin []
  (let [l1 is-fair
        (sample (flip 0.9))
        l2 coin
        (if is-fair (flip 0.5)
            (flip 0.95))]
    l3 (observe coin 1)
    l4 (observe coin 1)
    l5 (predict is-fair)) l6 )
```

$$S = [\![1, 6]\!] \times \{0, 1\}$$
$$f((1, 0)) := 1$$
$$F = \{(6, 0), (6, 1)\}$$

$$\delta((1, 0), (2, 0)) = 0, 1$$
$$\delta((1, 0), (2, 1)) = 0, 9$$

Exemple

```
(defquery coin []
  (let [l1 is-fair
        (sample (flip 0.9))
        l2 coin
        (if is-fair (flip 0.5)
            (flip 0.95))]
    l3 (observe coin 1)
    l4 (observe coin 1)
    l5 (predict is-fair)) l6 )
```

$$S = [\![1, 6]\!] \times \{0, 1\}$$
$$f((1, 0)) := 1$$
$$F = \{(6, 0), (6, 1)\}$$

$$\delta((1, 0), (2, 0)) = 0, 1$$
$$\delta((1, 0), (2, 1)) = 0, 9$$
$$\psi((4, 0)) = \psi((5, 0)) = 0, 95$$

Exemple

```
(defquery coin []
  (let [l1 is-fair
        (sample (flip 0.9))
        l2 coin
        (if is-fair (flip 0.5)
            (flip 0.95))]
    l3 (observe coin 1)
    l4 (observe coin 1)
    l5 (predict is-fair)) l6)
```

$S = [\![1, 6]\!] \times \{0, 1\}$
$f((1, 0)) := 1$
$F = \{(6, 0), (6, 1)\}$

$\delta((1, 0), (2, 0)) = 0, 1$
$\delta((1, 0), (2, 1)) = 0, 9$
$\psi((4, 0)) = \psi((5, 0)) = 0, 95$
$\psi((4, 1)) = \psi((5, 1)) = 0, 5$

Définition du problème d'inférence :

Rappel

$$\Pr(\theta|D,\mathcal{H}) = \frac{\Pr(D|\theta,\mathcal{H})\Pr(\theta|\mathcal{H})}{\Pr(D|\mathcal{H})}$$

Traces d'exécution de M : $\tau \in T = \bigsqcup_{1\leq n<\infty} S^n$. Espace mesuré $(T, \mathcal{G}, \nu)$.

Soit $n \geq 0$ et $\tau = s_0 \dots s_n \in T$. Ψ fonction de vraisemblance :

$$\Psi(\tau) = \prod_{0\leq i\leq n} \psi(s_i)$$

Δ densité de la probabilité à priori de la trace τ :

$$\Delta(\tau) = \prod_{0\leq i<n} [s_i \notin F] \cdot [s_n \in F] \cdot f(s_0) \cdot \prod_{0\leq i<n} \delta(s_i, s_{i+1})$$

Densité de probabilité postérieure $\Pi : T \to \mathbb{R}_+$

$$\Pi(\tau) = \frac{\Psi(\tau) \cdot \Delta(\tau)}{\int_T \Delta(\tau) \cdot \Psi(\tau)\, \nu(\mathrm{d}\tau)}$$

Définition (Approximation d'un PTS)

Un système de transition probabiliste est un 7-uplet $M = (S, \mathcal{F}, \mu, f, \delta, \psi, F)$ avec :

- (États) $(S, \mathcal{F}, \mu)$ est un espace mesuré tel que $\mu(S) < +\infty$;
- (États initiaux) $f : S \to \mathbb{R}_+$ est une fonction mesurable avec

$$\int_S f(s)\,\mu(\mathrm{d}s) = 1;$$

- (Transition) δ est une densité de transition sur $(S, \mathcal{F}, \mu)$;
- (Score) $\psi : S \to \mathbb{R}_+$ est une fonction mesurable ;
- (États finaux) $F \subseteq S$ est un sous-ensemble mesurable de S.

Définition (Approximation d'un PTS)

Un système de transition probabiliste approximatif est un 7-uplet $M_\theta = (S, \mathcal{F}, \mu, f_\theta, \delta_\theta, \psi, F)$ avec :

- (États) $(S, \mathcal{F}, \mu)$ est un espace mesuré tel que $\mu(S) < +\infty$;
- (États initiaux) $f_\theta : S \to \mathbb{R}_+$ est une fonction mesurable avec

$$\int_S f_\theta(s)\, \mu(\mathrm{d}s) = 1;$$

- (Transition) δ_θ est une densité de transition sur $(S, \mathcal{F}, \mu)$;
- (Score) $\psi : S \to \mathbb{R}_+$ est une fonction mesurable ;
- (États finaux) $F \subseteq S$ est un sous-ensemble mesurable de S.

Définition (Approximation d'un PTS)

Un système de transition probabiliste approximatif est un 7-uplet $M_\theta = (S, \mathcal{F}, \mu, f_\theta, \delta_\theta, \psi, F)$ avec :

- (États) $(S, \mathcal{F}, \mu)$ est un espace mesuré tel que $\mu(S) < +\infty$;
- (États initiaux) $f_\theta : S \to \mathbb{R}_+$ est une fonction mesurable avec

$$\int_S f_\theta(s)\, \mu(\mathrm{d}s) = 1;$$

- (Transition) δ_θ est une densité de transition sur $(S, \mathcal{F}, \mu)$;
- (Score) $\psi = \mathbf{1}$;
- (États finaux) $F \subseteq S$ est un sous-ensemble mesurable de S.

Définition (Approximation d'un PTS)

Un système de transition probabiliste approximatif est un 7-uplet $M_\theta = (S, \mathcal{F}, \mu, f_\theta, \delta_\theta, \psi, F)$ avec :

- (États) $(S, \mathcal{F}, \mu)$ est un espace mesuré tel que $\mu(S) < +\infty$;
- (États initiaux) $f_\theta : S \to \mathbb{R}_+$ est une fonction mesurable avec

$$\int_S f_\theta(s)\, \mu(\mathrm{d}s) = 1;$$

- (Transition) δ_θ est une densité de transition sur $(S, \mathcal{F}, \mu)$;
- (Score) $\psi = \mathbf{1}$;
- (États finaux) $F \subseteq S$ est un sous-ensemble mesurable de S.

Remarque

On suppose que $\Pi_\theta(\tau) = \Delta_\theta(\tau)$

Théorème (Forme générale de l'ELBO)

$$\mathcal{L}_{vi}(\Delta_\theta) = \int_T \Delta_\theta(\tau) \log\left(\frac{\Delta(\tau)\cdot\Psi(\tau)}{\Delta_\theta(\tau)}\right)\nu(\mathrm{d}\tau)$$

Théorème (Forme générale de l'ELBO)

$$\mathcal{L}_{vi}(\Delta_\theta) = \int_T \Delta_\theta(\tau) \log\left(\frac{\Delta(\tau)\cdot\Psi(\tau)}{\Delta_\theta(\tau)}\right)\nu(\mathrm{d}\tau) = \mathbb{E}_{\Delta_\theta}\left[\log\frac{\Delta(\tau)\cdot\Psi(\tau)}{\Delta_\theta(\tau)}\right]$$

Théorème (Forme simplifiée de l'ELBO)

$$\mathcal{L}_{vi} = \int f_\theta(s) \log \frac{f(s) \cdot \psi(s)}{f_\theta(s)} \mu(\mathrm{d}s) + \int \Delta_\theta(\tau) \left(\sum_{i=1}^{|\tau|-1} h_\theta(\tau_i) \right) \nu(\mathrm{d}\tau),$$

Avec :

$$h_\theta(s) = \int \delta_\theta(s, s') \log \frac{\delta(s, s') \cdot \psi(s')}{\delta_\theta(s, s')} \mu(\mathrm{d}s')$$

Théorème (Forme simplifiée de l'ELBO)

$$\mathcal{L}_{vi} = \mathbb{E}_{f_\theta(s)}\left[\log \frac{f(s)\cdot\psi(s)}{f_\theta(s)}\right] + \mathbb{E}_{\Delta_\theta(\tau)}\left[\sum_{i=1}^{|\tau|-1} h_\theta(\tau_i)\right]$$

Avec :

$$h_\theta(s) = \mathbb{E}_{\delta_\theta(s,s')}\left[\log \frac{\delta(s,s')\cdot\psi(s')}{\delta_\theta(s,s')}\right]$$

Conclusion

- Simplification du développement de modèles probabilistes ;

Conclusion

- Simplification du développement de modèles probabilistes ;
- Inférence “automatique” des programmes probabilistes ;

Conclusion

- Simplification du développement de modèles probabilistes ;
- Inférence “automatique” des programmes probabilistes ;
- Développement d’un nouvel algorithme d’inférence variationnelle ;

Conclusion

- Simplification du développement de modèles probabilistes ;
- Inférence “automatique” des programmes probabilistes ;
- Développement d’un nouvel algorithme d’inférence variationnelle ;
- Travail futur : implémentation et évaluation.

Questions

5. Questions
   - Complexité de l'inférence
   - Échantillonage
   - Descente stochastique de gradient
   - Théorie de la mesure
   - Simplification ELBO

Définition

Problème de décision BNI Étant donné un réseau bayésien $\mathcal{B}$ sur les variables $\mathcal{X}$, une variable $X \in \mathcal{X}$, et un événement $x \in val(X)$, décider si $\Pr(X = x) > 0$.

Théorème

BNI est $\mathcal{NP}$-Complet

BNI $\in \mathcal{NP}$.

- Choisir une assignation $\mathcal{A}$ des événements sur tout le réseau

Définition

Problème de décision BNI Étant donné un réseau bayésien $\mathcal{B}$ sur les variables $\mathcal{X}$, une variable $X \in \mathcal{X}$, et un événement $x \in val(X)$, décider si $\Pr(X = x) > 0$.

Théorème

BNI est $\mathcal{NP}$-Complet

BNI $\in \mathcal{NP}$.

- Choisir une assignation $\mathcal{A}$ des événements sur tout le réseau
- Si $\Pr(\mathcal{A}) > 0$ et $X = x$ dans $\mathcal{A}$, on a $\Pr(X = x) > 0$

□

BNI est $\mathcal{NP}$-difficile.

On réduit depuis 3-SAT. $\phi = C_1 \vee \ldots \vee C_m$ instance d'un problème sur les variables $q_1, \ldots, q_n$.

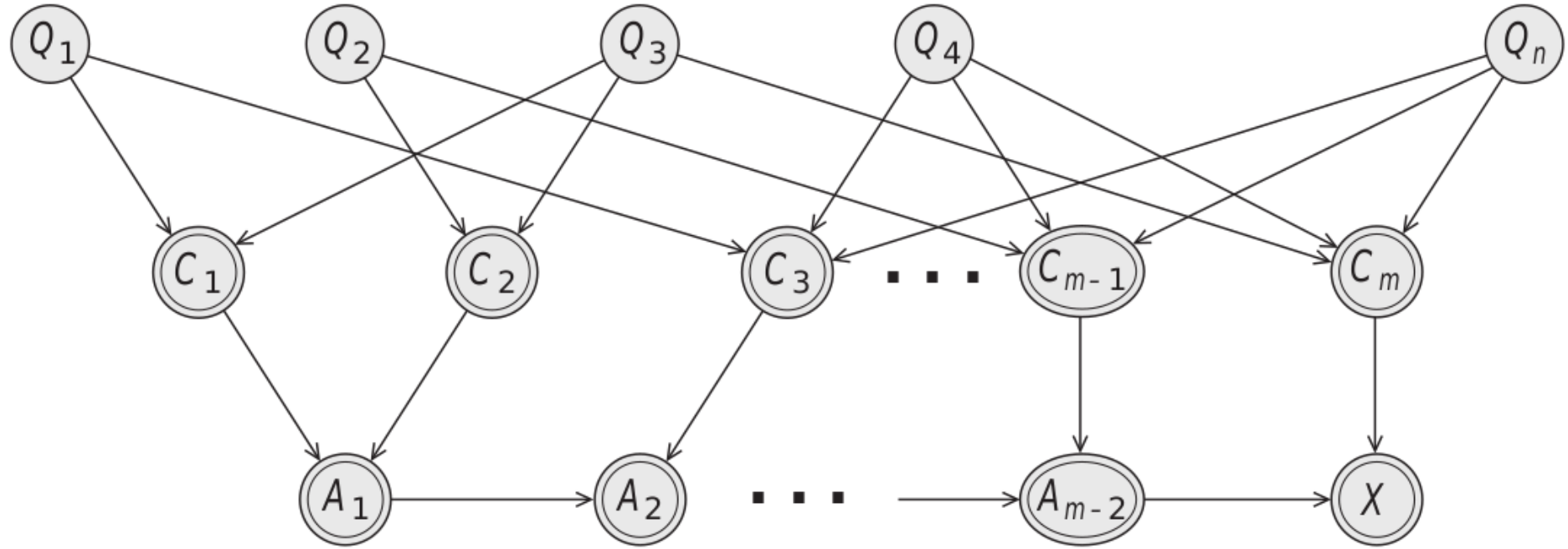


Figure issue de "Probabilistic Graphical Models : Principles and Techniques", Daphne Koller et Nir Friedman

Démonstration.

- $\Pr(Q_i = T) = 0.5$

Démonstration.

- $\Pr(Q_i = T) = 0.5$
- Arête entre Q_k et C_i si la variable q_k apparait dans C_i

Démonstration.

- $\Pr(Q_i = T) = 0.5$
- Arête entre Q_k et C_i si la variable q_k apparait dans C_i
- Table de C_i : se comporte comme la clause en fonction des variables (table booléenne, pas probabiliste)

Démonstration.

- $\Pr(Q_i = T) = 0.5$
- Arête entre Q_k et C_i si la variable q_k apparait dans C_i
- Table de C_i : se comporte comme la clause en fonction des variables (table booléenne, pas probabiliste)
- Les A_i sont des portes logiques ET (pour que la taille du problème reste polynomiale en l'entrée)

Démonstration.

- $\Pr(Q_i = T) = 0.5$
- Arête entre Q_k et C_i si la variable q_k apparait dans C_i
- Table de C_i : se comporte comme la clause en fonction des variables (table booléenne, pas probabiliste)
- Les A_i sont des portes logiques ET (pour que la taille du problème reste polynomiale en l'entrée)
- X est le ET de C_m et A_{m-2}, sera la bonne variable, ie $X = 1 \Leftrightarrow$ toutes les clauses sont satisfaites

Démonstration.

- $\Pr(Q_i = T) = 0.5$
- Arête entre Q_k et C_i si la variable q_k apparait dans C_i
- Table de C_i : se comporte comme la clause en fonction des variables (table booléenne, pas probabiliste)
- Les A_i sont des portes logiques ET (pour que la taille du problème reste polynomiale en l'entrée)
- X est le ET de C_m et A_{m-2}, sera la bonne variable, ie $X = 1 \Leftrightarrow$ toutes les clauses sont satisfaites
- $\Pr(X = T) = \Pr(X = T|q_i)\Pr(q_i) = \Pr(X = T|q_i)2^{-n}$

Démonstration.

- $\Pr(Q_i = T) = 0.5$
- Arête entre Q_k et C_i si la variable q_k apparait dans C_i
- Table de C_i : se comporte comme la clause en fonction des variables (table booléenne, pas probabiliste)
- Les A_i sont des portes logiques ET (pour que la taille du problème reste polynomiale en l'entrée)
- X est le ET de C_m et A_{m-2}, sera la bonne variable, ie $X = 1 \Leftrightarrow$ toutes les clauses sont satisfaites
- $\Pr(X = T) = \Pr(X = T|q_i)\Pr(q_i) = \Pr(X = T|q_i)2^{-n}$
- Donc $\Pr(X = T) > 0 \Longleftrightarrow \phi$ est satisfiable

Inférence par échantillonage

Échantillonage

Estimation de $\mathbb{E}_{p(x)}[f(x)]$ par $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ avec $x_1, \ldots, x_n \sim p(x)$

Inférence par échantillonage

Échantillonage

Estimation de $\mathbb{E}_{p(x)}[f(x)]$ par $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ avec $x_1, \ldots, x_n \sim p(x)$

Méthode du rejet : estimation de $\Pr(R = T | W = T, S = T)$

Entrée: Nombre N d'échantillons

```
1 rain := 0
2 n := 0
3 for i = 1 to N do
4 |   choose w, s, r, c ~ Pr(W, S, R, C)
5 |   if w and s then
6 |   |   n := n + 1
7 |   |   if r then rain := rain + 1
8 return rain/n
```

Inférence dans les programmes probabilistes

Idée de l'algorithme du rejet :

- Interpréter le programme jusqu'à rencontrer `sample`, `observe`, `predict`
- `sample dist` : tirer un échantillon
- `observe dist value` : tirer un échantillon v. Si $v \neq value$, recommencer
- `predict expr` : garder en mémoire la valeur de `expr`

Input: Number N of steps
1 Choose $\theta \in \Theta$ randomly
2 **for** $i = 1$ *to* N **do**
3 $\quad \theta := \theta + \eta(i) \cdot \nabla_\theta \mathcal{L}_{vi}(\Delta_\theta)$
4 **return** θ

Algorithm 1: Montée stochastique de gradient

Les conditions de Robbins-Monro assurent une convergence vers un maximum :

$$\sum_{i=1}^{\infty} \eta(i) = \infty \qquad \sum_{i=1}^{\infty} \eta(i)^2 < \infty$$

Définition (σ-algèbre)

Soit X un ensemble. $\Sigma \subseteq \mathcal{P}(X)$ est une σ-algèbre sur X si :

- $\emptyset \in \Sigma$,
- $\forall B \in \Sigma$, $\overline{B} \in \Sigma$,
- si $\forall n \in \mathbb{N}, B_n \in \Sigma$, alors $\bigcup_{n \in \mathbb{N}} B_n \in \Sigma$.

Remarque

Dans la suite, X est un ensemble, et Σ est une σ-algèbre sur X. (X, Σ) est un espace mesurable.

Définition (Mesure)

$\mu : \Sigma \to \overline{\mathbb{R}}_+$ est une mesure si

- $\forall E \in \Sigma,\ \mu(E) \geq 0$,
- $\mu(\emptyset) = 0$,
- si $(E_i)_{i \in I}$ est une suite d'élément disjoint deux à deux, $\mu(\bigcup_{i \in I} E_i) = \sum_{i \in I} \mu(E_i)$.

Définition

Une mesure μ sur (X, Σ) est *σ-finie* si $\mu(X) < +\infty$.

Définition (Mesure de probabilité)

Une mesure de probabilité μ est une mesure telle que $\mu(X) = 1$.

Exemple (Mesure de Dirac)

Soit $x \in X$, $A \in \Sigma$. On défini la mesure de Dirac au point x, notée δ_x , par la mesure suivante : $\delta_x(A) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{si } x \notin A \end{cases}$

Définition (Espace mesuré)

Un espace mesuré est un triplet (X, Σ, μ), où X est un ensemble, Σ est une σ-algèbre sur X et μ est une mesure sur (X, Σ).

Définition (Ensemble mesurable)

Soit $A \subseteq X$, A est mesurable si $A \in \Sigma$.

Définition (Fonction mesurable)

Soit (Y, T) un espace mesurable. $f : X \to Y$ est mesurable si :

$$\forall E \in T, \{x \in X \mid f(x) \in E\} \in \Sigma$$

- Utilisation des A_θ, B_θ
- Simplification vers les D_θ et E_θ

Avoir le rapport sous la main ? Tellement plus simple...