Cours IFT6266,

Apprentissage Bayesien

Le point de vue Bayesien sur l'apprentissage est le suivant. À n'importe quel moment l'apprenant a une incertitude sur la réalité qui s'exprime comme une distribution de probabilité dans l'espace des modèles, i.e., des interprétations possibles de cette réalité. Cette distribution de probabilité doit être interprétée comme une **croyance**, et non pas comme la limite d'une fréquence. Toute incertitude est représentée par une distribution. Quand on observe des exemples, cela nous permet de réviser notre incertitude sur notre modèle, généralement en rendant cette distribution plus piquée. Dans le cas où les modèles que nous considérons sont représentés par un paramètre θ, le résultat de l'apprentissage n'est donc pas une valeur particulière pour θ mais une distribution *a posteriori* sur θ. Cependant, il faut partir d'une distribution *a priori* sur θ, qui représente les modèles que l'ont croit possibles avant de voir les exemples.

Le théorème de Bayes permet de relier distribution *a priori* $P(\theta)$ et *a posteriori* $P(\theta|D)$:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}.$$

La quantité $P(D|\theta)$ est appelée **vraisemblance** (likelihood). Notons que l'on s'intéresse à cette valeur et à $P(D|\theta)$ simultanément pour TOUTES les valeurs de θ, ce qui pose certains défis calculatoires. Le dénominateur, qui égale $P(D)$, est aussi appelé **l'évidence** et il sert uniquement à normaliser. Il est donc inutile si notre but est seulement de trouver le θ le plus probable, mais il est essentiel si on veut faire n'importe quoi d'autre (comme par exemple prédire un Y à partir d'un X).

La conclusion de l'inférence Bayesienne doit être comprise comme étant dépendente de la distribution *a priori* qui a été choisie. Si on part avec un "mauvais" a priori, on obtient de mauvais résultats. En particulier, si le "vrai" modèle générateur de D n'est pas inclus dans le modèle (i.e. $\nexists\theta$ tel que $P(D|\theta)$ est le vrai générateur), alors on ne peut rien conclure.

Par contre il y a des avantages à l'inférence Bayesienne: elle tient automatiquement compte de l'incertitude due au nombre d'exemples dans les intervalles de confiance qu'elle nous donne quand aux prédictions faites.

De plus, elle est théoriquement à l'abris de l'overfitting (on va automatiquement se retrouver avec une fonction de décision plus "simple" quand on a peu d'exemples).

C'est "théoriquement" car cela dépend de la justesse de la loi a priori, mais en pratique c'est une bonne manière d'éviter l'overfitting.

En résumé:

- On considère la quantité à estimer (e.g. paramètre θ) comme une v.a. qui caractérise un modèle $P(D|\theta)$ des données D.
- On suppose qu'on a une croyance ou connaissance à priori sur θ, exprimée par la loi a priori $P(\theta)$.
- On suppose que la loi $P(D|\theta)$ est connue quand θ est connu.
- On applique alors le théorème de Bayes et les lois des probabilités pour toute inférence. En général:
  - Si on doit **prédire** la loi d'une v.a Y impliquée dans le modèle étant (optionnellement) donnée une information supplémentaire X, alors on prend "simplement" $P(Y = y|D, X)$ (ce qui implique une somme ou une intégrale sur tous les θ possibles).
  - Si on doit **choisir** une valeur pour une v.a. Y , avec une fonction de coût $C(choix, Y)$ alors on choisit $\text{argmin}_{choix} E[C(choix, Y)|D, X)]$ (ce qui implique encore une somme ou une intégrale sur tous les θ possibles).

  Exemples:

  - Choisir le θ le plus probable: $\text{argmax}_\theta P(\theta|D) = \text{argmax}_\theta P(D|\theta)P(\theta)$.
  - Régression Bayesienne: $E[Y|D, X] = \int E[Y|X, \theta]P(\theta|D)d\theta = \frac{\int E[Y|X,\theta]P(D|\theta)P(\theta)d\theta}{\int P(D|\theta)P(\theta)}$.
- En général on doit faire des intégrales pas faciles et beaucoup de techniques ont été proposées pour les approximer.

Il y a un cas particulier intéressant. Si notre "décision" consiste à choisir θ, alors on obtient la solution appelée MAP (maximum a posteriori), qui est équivalente à la solution d'un problème classique (non-Bayesien) d'apprentissage, avec un terme de régularisation qui représente notre a priori:

$$\begin{aligned}
\text{argmax}_\theta P(\theta|D) &= \text{argmax}_\theta \frac{P(D|\theta)P(\theta)}{P(D)} \\
&= \text{argmax}_\theta P(D|\theta)P(\theta) \\
&= \text{argmax}_\theta \log P(D|\theta) + \log P(\theta) \\
&= \text{argmin}_\theta \ -\log P(D|\theta) - \log P(\theta) \\
&= \text{argmin}_\theta \ \mathit{cout} \ + \ \lambda\Omega(\theta)
\end{aligned}$$

ce qui montre que dans l'interprétation Bayesienne, le coût à minimiser est moins la log-vraisemblance $-\log P(D|\theta)$, et le terme de régularisation est moins le log-prior $-\log P(\theta)$ (possiblement à une constante près).

Il y a aussi des cas (très) particuliers de l'inférence Bayesienne dans lesquels on peut faire tous les calculs de manière analytique. En particulier, si on prend une distribution a priori qui est dite **conjuguée** avec la vraisemblance, cela veut dire que la distribution a posteriori se retrouve dans la même famille de distributions que la distribution a priori. Par exemple, si la loi a priori est Normale, et que la vraisemblance est normale (par exemple comme pour la régression linéaire), alors la loi a posteriori est aussi Normale. Comme ces lois sur θ sont paramétriques, on peut les spécifier de manière finie et exacte avec leurs paramètres (e.g., moyenne, matrice de covariance). Dans beaucoup de ces cas, on peut faire analytiquement les calculs qui nous donnent les paramètres a posteriori étant donnés les paramètres a priori et les données d'apprentissage.

Considérons le cas spécial de la régression linéaire avec bruit connu. On a au départ une distribution à priori sur les paramètres w et à l'arrivée une distribution à postériori sur ces mêmes paramètres, après avoir vu les données $D = \{(x_t, y_t)\}_{t=1}^n$, $x_t \in \mathbb{R}^d$, $y \in \mathbb{R}$. On suppose $Y|X = x$ normal de variance σ^2 (connue pour simplifier) et d'espérance $w'x$. Soit $p(w) = \mathcal{N}(w|m_0, S_0)$ la loi à priori sur w, et notons $\mathbf{X}$ la matrice dont les rangées sont les x_t et $\mathbf{Y}$ le vecteur colonne dont les éléments sont les y_t. Le résultat a posteriori est:

$$p(w|\mathbf{X}, \mathbf{Y}) = \mathcal{N}(w|m_n, S_n)$$

avec

$$\begin{aligned} m_n &= S_n(S_0^{-1}m_0 + \mathbf{X}'\mathbf{Y}/\sigma^2) \\ S_n^{-1} &= S_0^{-1} + \mathbf{X}'\mathbf{X}/\sigma^2. \end{aligned}$$

Pour démontrer ce résultat, on exploite le fait qu'une densité de la forme $ce^{a'x+x'Bx}$ (avec $-B$ positive semi-définie) est forcément Normale, et il suffit de compléter le carré pour mettre le polynôme dans l'exponentielle sous la forme $-0.5(\mu - x)'\Sigma^{-1}(\mu - x)$ afin d'identifier la moyenne μ et la covariance Σ de cette Normale. Notons qu'il faut absolument que la matrice Σ obtenue soit positive définie.

Contrairement à la régression linéaire ordinaire, on obtient non seulement une prédiction mais aussi une incertitude sur les prédictions qui sera d'autant plus petite qu'il y a beaucoup d'exemples. On peut ainsi utiliser la formule ci-haut pour obtenir la distribution à postériori sur Y pour un nouvel exemple $X = x$, i.e., obtenir une formule pour $p(Y|X = x, \mathbf{X}, \mathbf{Y})$. Remarquez comme les paramètres w ont été intégrés, et donc n'apparaissent pas dans le résultat.