Apprentissage de modèles probabilistes

Chapter 20

Plan

♢ Apprentissage Bayésien

♢ Apprentissage du Maximum *à posteriori* et de la vraisemblance (maximum likelihood–ML)

♢ Apprentissage des réseaux Bayésiens

– apprentissage des paramètres de la vraisemblance (ML) avec données complètes

– Régression linéaiare

Apprentissage Bayesien

L'apprentissage est vu ici comme une mise à jour Bayésienne de la distribution de probabilité sur l'espace des hypothèses

H est la variable hypothèse dont les valeurs sont $h_1, h_2, \ldots$, et le prior $\mathbf{P}(H)$

La $j$ième observation d_j donne l'instantiation de la variable aléatoire D_j
Données d'entraînement sont $\mathbf{d} = d_1, \ldots, d_N$

Selon la règle de Bayès, chaque hypothèse a comme probabilité à posteriori :

$$P(h_i|\mathbf{d}) = \alpha P(\mathbf{d}|h_i)P(h_i)$$

où $P(\mathbf{d}|h_i)$est appelée la vraisemblance ou le likelihood

La prédiction d'une variable inconnue X utilise la moyenne pondérée de la vraisemblance sur les différentes hypothèses :

$$\mathbf{P}(X|\mathbf{d}) = \Sigma_i \, \mathbf{P}(X|\mathbf{d}, h_i)P(h_i|\mathbf{d}) = \Sigma_i \, \mathbf{P}(X|h_i)P(h_i|\mathbf{d})$$

Apprentissage Bayesien

La prédiction

$$\mathbf{P}(X|\mathbf{d}) = \Sigma_i \, \mathbf{P}(X|h_i) P(h_i|\mathbf{d})$$

montre que les hypothèses sont des "intermédiaires" entre les données brutes et les prédictions.

Les quantités clés dans l'approche bayésienne sont donc
– la probabilité à priori, $P(h_i)$
– la vraisemblance de la donnée selon chaque hypothèse, $P(d|h_i)$

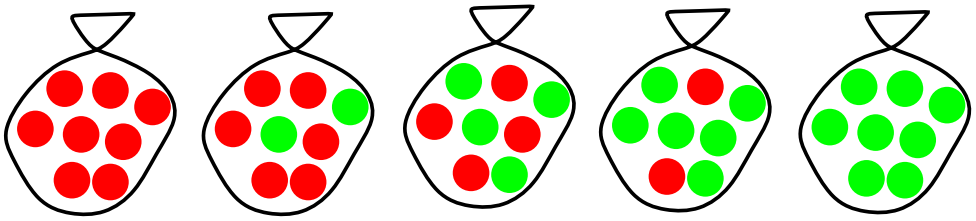
Exemple si l'on $h_1, \ldots, h_5$ et que la distribution à priori est : $\langle 0.1, 0.2, 0.4, 0.2, 0.1 \rangle$. Alors la probabilité des données se calcule en supposant que les observations sont distribuées de manière indépendante et identique (i.i.d), de sorte que :

$$P(d|h_i) = \Pi_j P(d_j|h_i)$$

Exemple

Supposons que nous ayons 5 genres de sacs de bonbons :

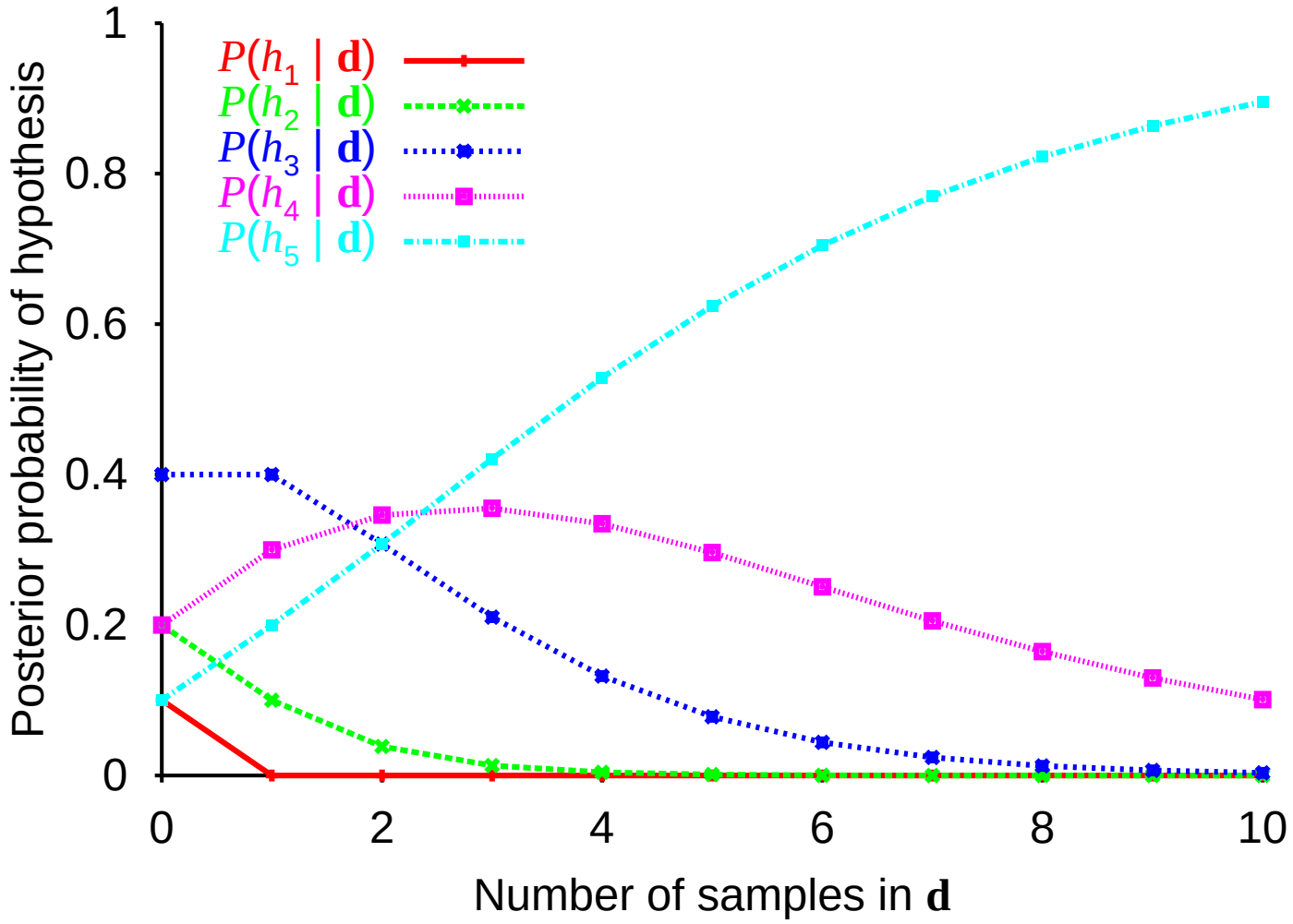
10% est h_1 : 100% cherry candies
- 20% est h_2 : 75% cherry candies + 25% lime candies
- 40% est h_3 : 50% cherry candies + 50% lime candies
- 20% est h_4 : 25% cherry candies + 75% lime candies
- 10% est h_5 : 100% lime candies



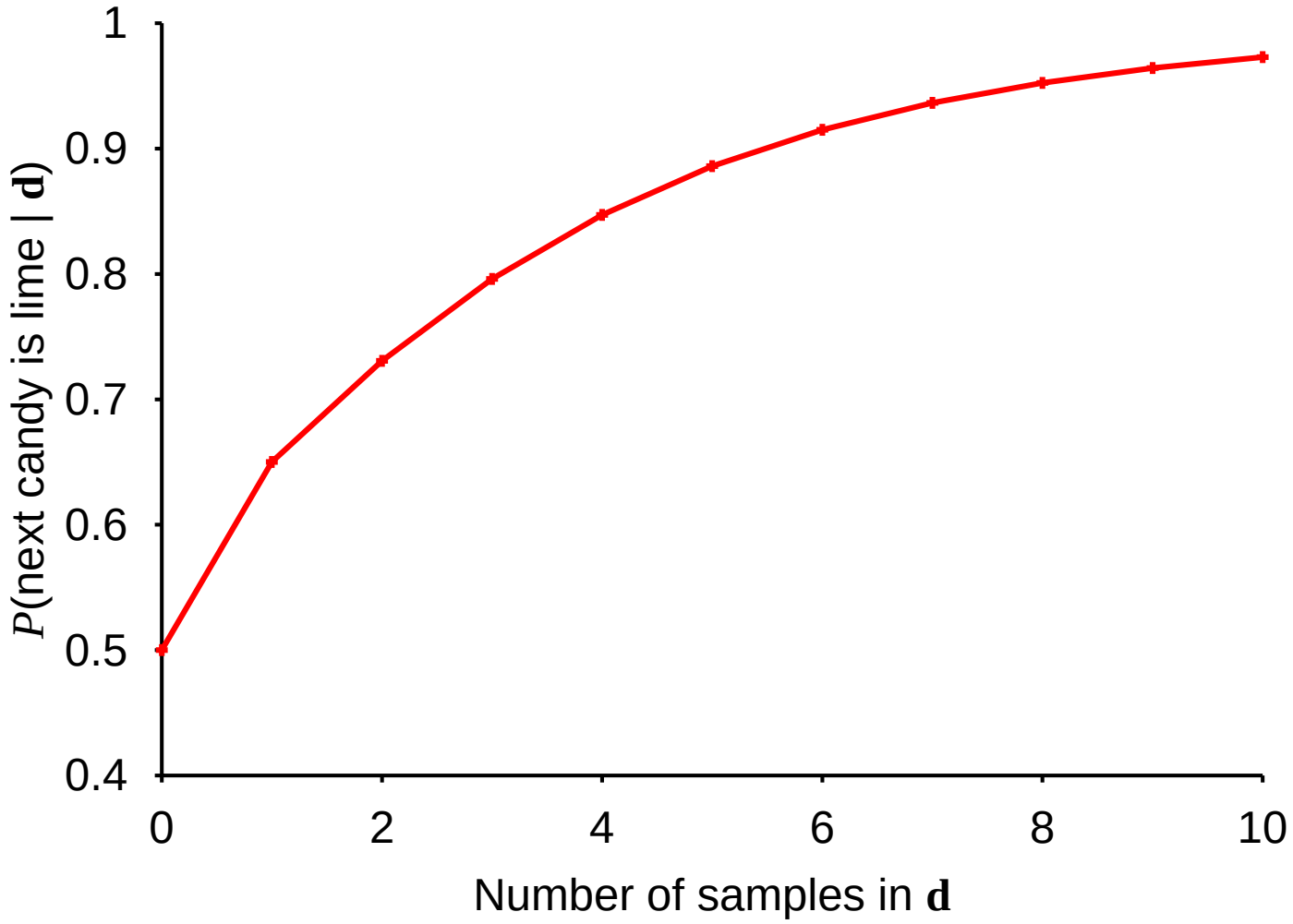
On observe alors des bonbons tirés d'un sac donné : 

Quel est le genre de sac visé (quelle est l'hypothèse h_i ?) Quelle pourrait être alors le genre de bonbon qu'on peut tirer au 11ème coup ?

Probabilite a posteriori des hypotheses



Probabilite pour la prediction



Approximation : le MAP

Une sommation sur un large espace d'hypothèses est souvent intraitable (i.e., intractable)
(e.g., 18,446,744,073,709,551,616 Fonction booléennes de 6 attributes)

L'apprentissage Maximum a posteriori (MAP) : ça consiste à choisir h_{MAP} maximisant $P(h_i|\mathbf{d})$

I.e., maximiser $P(\mathbf{d}|h_i)P(h_i)$ ou $\log P(\mathbf{d}|h_i) + \log P(h_i)$

Les Log terms peuvent être vus comme
des bits pour encoder les données étant donné les hypothèses + des bits pour encoder les hypothèses
Cette idée est à la base de l'apprentissage utilisant la longueur de description minimale, (i.e., minimum description length (MDL) learning

MAP = exprime la simplicité en assignant des probabilités plus élevées aux hypothèses consistantes (i.e., simplest consistent hypothesis (cf. science))

Approximation : le ML

Pour de grands ensembles de données, le prior devient non-pertinent

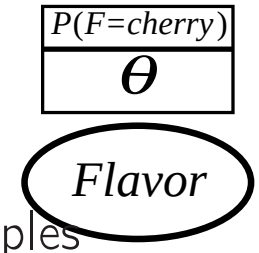
Dès lors il convient d'utiliser l'apprentissage utilisant la vraisemblance, soit Maximum likelihood (ML) learning : ça consiste à choisir h_{ML} qui maximize $P(\mathbf{d}|h_i)$

I.e., simplement obtenir le meilleur "fit" des données
ML est donc identique au MAP pour des prior uniformes

ML est la méthode d'apprentissage statique non Bayésienne standard

RB : Apprentissage de parametres

On a un sac de bonbons venant d'un nouveau fabricant ; on voudrait alors savoir : la fraction θ de bonbons du type cherry ?



Tout θ est possible : on a donc un continuum d'hypothèses h_θ
θ est un paramètre pour cette famille de modèles binomiales simples

Supposons qu'on ouvre N bonbons, c cherries et $\ell = N - c$ limes
Ces observations sont i.i.d. (independent, identically distributed), et par conséquent

$$P(\mathbf{d}|h_\theta) = \prod_{j=1}^{N} P(d_j|h_\theta) = \theta^c \cdot (1-\theta)^\ell$$

RB : Apprentissage de parametres (suite)

Maximiser ceci relativement à θ—plus facile si on passe par le log-likelihood :

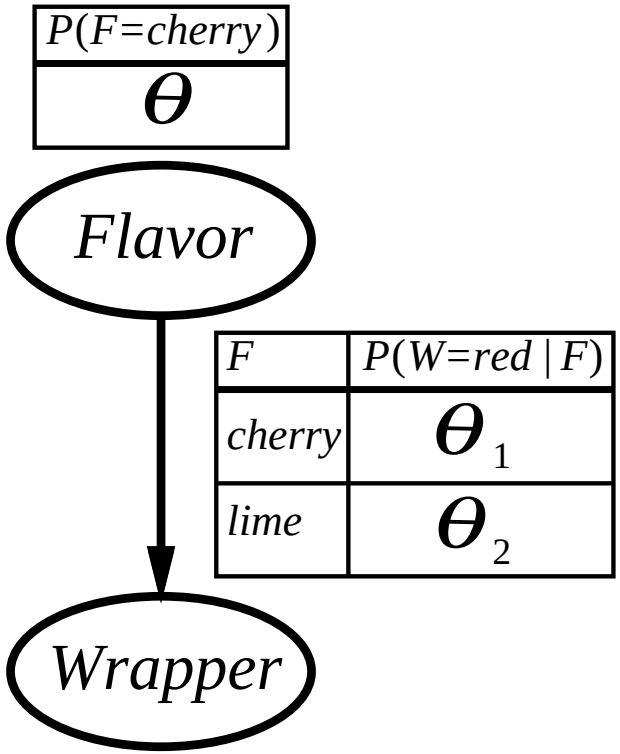
$$L(\mathbf{d}|h_\theta) = \log P(\mathbf{d}|h_\theta) = \sum_{j=1}^{N} \log P(d_j|h_\theta) = c \log\theta + \ell \log(1-\theta)$$

$$\frac{dL(\mathbf{d}|h_\theta)}{d\theta} = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \qquad \Rightarrow \qquad \theta = \frac{c}{c+\ell} = \frac{c}{N}$$

Semble conforme à l'intuition mais pose problème pour $N = 0$.

Multiple parametres

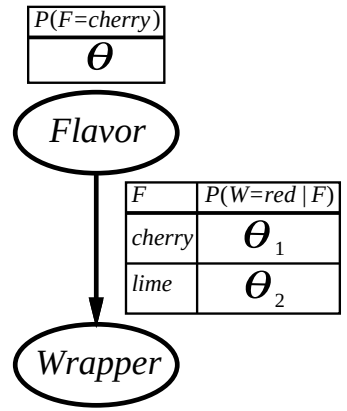
L'emballage Rouge/vert (Red/Green) va dépendre probablistiquement de l'arôme :



Multiple parametres

La vraisemblance (Likelihood) pour, e.g., cherry candy utilisant le papier green :

$$\begin{aligned} & P(F = cherry, W = green | h_{\theta,\theta_1,\theta_2}) \\ &= P(F = cherry | h_{\theta,\theta_1,\theta_2}) P(W = green | F = cherry, h_{\theta,\theta_1,\theta_2}) \\ &= \theta \cdot (1 - \theta_1) \end{aligned}$$



N candies, r_c red-wrapped cherry candies, etc. :

$$P(\mathbf{d} | h_{\theta,\theta_1,\theta_2}) = \theta^c (1-\theta)^\ell \cdot \theta_1^{r_c} (1-\theta_1)^{g_c} \cdot \theta_2^{r_\ell} (1-\theta_2)^{g_\ell}$$

$$\begin{aligned} L &= [c \log \theta + \ell \log(1-\theta)] \\ &+ [r_c \log \theta_1 + g_c \log(1-\theta_1)] \\ &+ [r_\ell \log \theta_2 + g_\ell \log(1-\theta_2)] \end{aligned}$$

Multiple parametres (suite)

Les dérivées de L par rapport à chacun des paramètres donnent alors :

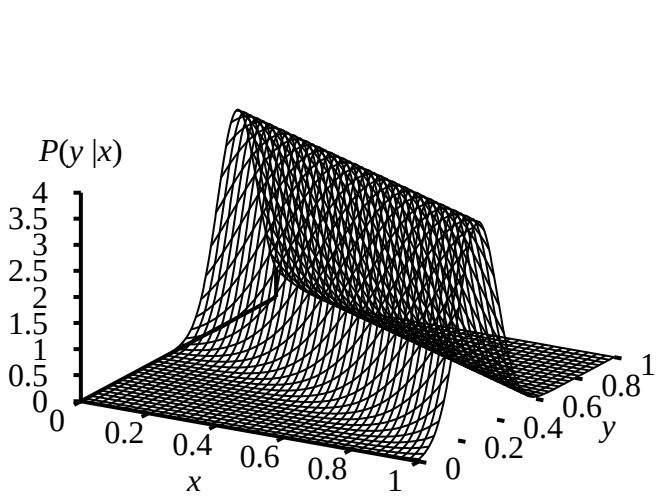
$$\frac{\partial L}{\partial \theta} = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \qquad \Rightarrow \quad \theta = \frac{c}{c+\ell}$$

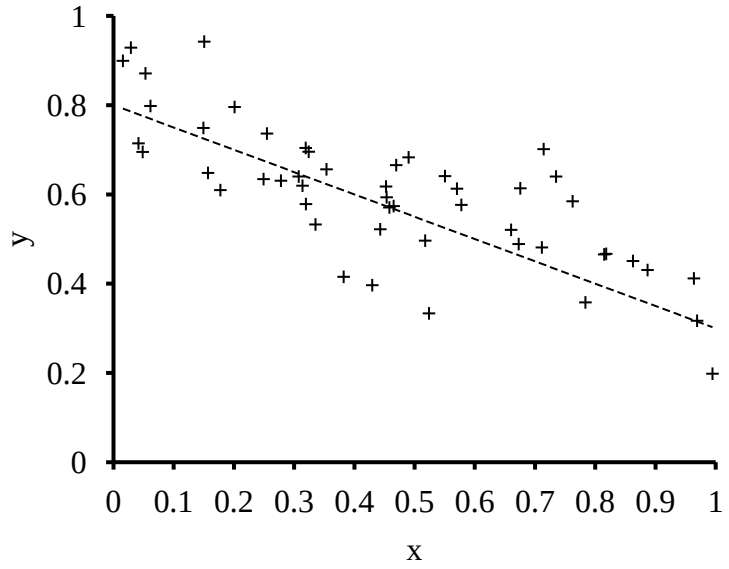
$$\frac{\partial L}{\partial \theta_1} = \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 \qquad \Rightarrow \quad \theta_1 = \frac{r_c}{r_c+g_c}$$

$$\frac{\partial L}{\partial \theta_2} = \frac{r_\ell}{\theta_2} - \frac{g_\ell}{1-\theta_2} = 0 \qquad \Rightarrow \quad \theta_2 = \frac{r_\ell}{r_\ell+g_\ell}$$

En utilisant des données complètes, **le problème de l'apprentissage des paramètres par ML pour un RB se décompose en problèmes séparés : un pour chaque paramètre**

Exemple : le modele Gaussien lineaire





Maximizing $P(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-(\theta_1 x+\theta_2))^2}{2\sigma^2}}$ w.r.t. θ_1, θ_2

= minimizing $E = \sum_{j=1}^{N} (y_j - (\theta_1 x_j + \theta_2))^2$

Autrement dit, minimiser la somme des carrés des erreurs donne le modèle linéaire de vraisemblance maximale, pourvu que les données soient générées avec un bruit gaussien de variance fixe.

Sommaire

L'apprentissage bayésien donne les meilleures prédictions mais il est intraitable sur machine

L'apprentissage MAP équilibre complexité et précision sur les données d'entraînement

L'apprentissage ML suppose un prior uniforme, il est bon en particulier pour les larges data sets. Il consiste à

1. choisir une famille de modèles paramétrisée pour décrire les données
ça requiert un substantiel discernment et parfois de nouveaux modèles

2. formuler le ML comme une fonction de paramètres
pourrait nécessiter de sommer sur des variables cachées

3. dériver le ML relativement à chacun des paramètres

4. trouver les valeurs des paramètres en égalisant les dérivées à zero
peut être difficile/impossible ; faire appel aux tech. d'optimisation modernes