Lezione 6

Bayesian Learning

Martedì, 16 Novembre 2004

Giuseppe Manco

Readings:
Sections 6.1-6.5, Mitchell
Chapter 2, Bishop
Chapter 4, Hand-Mannila-Smith

Concetti probabilistici, apprendimento probabilistico

- **Concetti probabilistici**
  - **Il concetto da apprendere è una funzione *c*: $X \rightarrow [0, 1]$**
  - ***c*(*x*), il valore target, denota la probabilità che l'etichetta 1 (*True*) sia assegnata a *x***
  - **Sostanzialmente, quello che abbiamo ottenuto finora**
- **Probabilistic (i.e., Bayesian) Learning**
  - **Utilizzo di un criterio probabilistico nella selezione di un'ipotesi *h***
    - **Esempio: l'ipotesi "più probabile" *considerando D*: MAP hypothesis**
    - **Esempio: *h* per cui *D* è "più probabile": max likelihood (ML) hypothesis**
  - **NB: *h* può essere una qualsiasi funzione**

Alcune definizioni di base

- **Spazio degli eventi (Ω): Dominio di una variabile casuale X**
- **Misura di probabilità $P(\bullet)$**
  - **P, è una misura su Ω**
  - **$P(X = x \in \Omega)$ è una misura della fiducia in $X = x$**

- **Assiomi di Kolmogorov**
  - **1. $\forall x \in \Omega \,.\, 0 \leq P(X = x) \leq 1$**
  - **2. $P(\Omega) \equiv \sum_{x \in \Omega} P(X = x) = 1$**
  - **3. $\forall X_1, X_2, \ldots \ni i \neq j \Rightarrow X_i \wedge X_j = \varnothing$.**

$$P\left(\bigcup_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} P(X_i)$$

- **Probabilità congiunta: $P(X_1 \wedge X_2) \equiv$ dell'evento $X_1 \wedge X_2$**
- **indipendenza: $P(X_1 \wedge X_2) = P(X_1) \bullet P(X_2)$**

Il teorema di Bayes

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} = \frac{P(h \wedge D)}{P(D)}$$

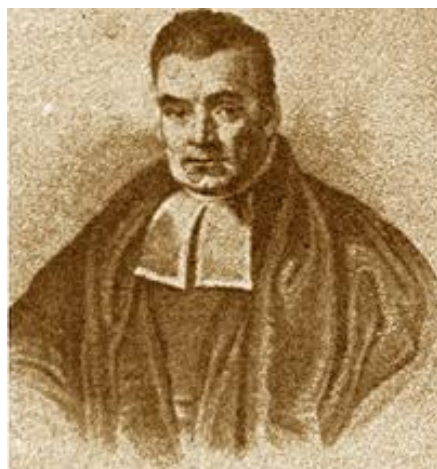
- $P(h)$ ≡ Probabilità a priori dell'ipotesi h
  - Misura la credenza iniziale indipendentemente da qualsiasi informazione (e quindi a priori)
- $P(D)$ ≡ Prior dell'insieme D
  - Misura la probabilità dell'insieme D (i.e., expresses D)
- $P(h \mid D)$ ≡ Probabilità di h dato D
  - | denota condizionamento - P(h | D) is probabilità condizionale (*a* posteriori)
- $P(D \mid h)$ ≡ Probabilità di D dato h
  - Probabilità di osservare D sapendo che vale h (modello "generativo")
- $P(h \wedge D)$ ≡ Probabilità congiunta di h e D

Da Bayes "Essay towards solving a problem in the doctrine of chances" (1763)

Thomas Bayes

Born: 1702 in London, England

Died: 1761 in Tunbridge Wells, Kent, England



La scelta delle ipotesi

- **Teorema di Bayes**

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} = \frac{P(h \wedge D)}{P(D)}$$

- **Ipotesi MAP**
  - Si vuole l'ipotesi più probabile sullo specifico training set
  - $\arg\max_{x \in \Omega}[f(x)]$ ≡ il valore di *x* nello spazio Ω che esibisce il più alto *f*(*x*)
  - Maximum *a posteriori* hypothesis, h_{MAP}

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h \mid D) \\ &= \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D \mid h)P(h) \end{aligned}$$

- **Ipotesi ML**
  - Assumiamo $p(h_i) = p(h_j)$ (tutte le ipotesi sono equalmente probabili)
  - Si sceglie l'ipotesi che spiega meglio I dati, h_{ML}

$$h_{ML} = \arg\max_{h_i \in H} P(D \mid h_i)$$

Esempio: Moneta bilanciata [1]

- **Lancio della moneta**
  - Spazio: Ω = {*Head*, *Tail*}
  - Scenario: la moneta è bilanciata o sbilanciata al 60% in favore di *Head*
    - $h_1 \equiv$ bilanciata: $P(Head) = 0.5$
    - $h_2 \equiv$ 60% bias : $P(Head) = 0.6$
  - Obiettivo: decidere tra l'ipotesi di default (null) e l'alternativa
- **Distrivuzione a Priori**
  - $P(h_1) = 0.75$, $P(h_2) = 0.25$
  - Riflette le credenze iniziali su H
  - L'apprendimento è revisione delle credenze
- **Evidenze**
  - $d \equiv$ singolo lancio, viene *Head*
  - D: Cosa crediamo adesso?
  - R: Calcoliamo $P(d) = P(d \mid h_1)\, P(h_1) + P(d \mid h_2)\, P(h_2)$

Esempio: Moneta bilanciata [2]

- Inferenza Bayesiana: Calcoliamo $P(d) = P(d \mid h_1)\, P(h_1) + P(d \mid h_2)\, P(h_2)$
  - $P(Head) = 0.5 \bullet 0.75 + 0.6 \bullet 0.25 = 0.375 + 0.15 = 0.525$
  - Questa è la probabilità dell'osservazione $d = Head$
- Apprendimento bayesiano
  - In base al teorema di Bayes
    - $P(h_1 \mid d) = P(d \mid h_1)\, P(h_1) / P(d) = 0.375 / 0.525 = 0.714$
    - $P(h_2 \mid d) = P(d \mid h_2)\, P(h_2) / P(d) = 0.15 / 0.525 = 0.286$
    - *Le credenze sono state spostate verso* h_1
    - MAP: crediamo ancora che la moneta sia bilanciata
  - Approccio ML (assumiamo priors identici)
    - Le credenze sono revisionate a partire da 0.5
    - C'è più sbilanciamento a favore di h_1
- Ulteriore evidenza: Sequenza D di 100 lanci con 70 heads e 30 tails
  - $P(D) = (0.5)^{70} \bullet (0.5)^{30} \bullet 0.75 + (0.6)^{70} \bullet (0.4)^{30} \bullet 0.25$
  - Ora $P(h_1 \mid D) << P(h_2 \mid D)$

Stima di densità

- **Obiettivo principale: stimare *P*(*D* | *h*)**
  - Grazie al teorema di Bayes, possiamo ottenere la probabilità “a posteriori”
- **Tre approcci**
  - **Metodi parametrici**
    - **Si assume una forma funzionale per le densità**
    - **Si stimano i parametri di tali forme funzionali**
  - **Metodi nonparametrici**
    - **La forma funzionale è determinata dai dati**
  - **Mixture models**
    - **Combinazioni di molte possibili forme funzionali**
    - **Neural Networks**

Esempio: Maximum Likelihood Estimation

- **Dati M parametri**
- **Si trovino i valori più probabili**
  - **Massimizzando la probabilità congiunta**
    - **Assumendo N istanze indipendenti**

$$\boldsymbol{\theta} = [\theta_1, \ldots, \theta_M]$$

$$L(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{i=1}^{N} p(\mathbf{x}_i \mid \boldsymbol{\theta})$$

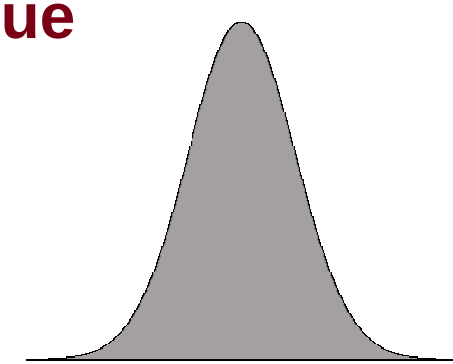
  - **Minimizzando l'errore corrispondente**

$$E = -\ln L(\mathbf{x}_1, \ldots, \mathbf{x}_N) = -\sum_{i=1}^{N} \ln p(\mathbf{x}_i \mid \boldsymbol{\theta})$$

Il caso di dati gaussiani

- **Assunzione tipica: gli attributi numerici hanno una distribuzione normale (*Gaussiana)***
- **La densità di probabilità è definita da due parametri**
  - **densità *f(x):***

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Dati gaussiani e MLE

- **Stimiamo la verosimiglianza**

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/(2\sigma^2)} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)\sum_{i=1}^{n}(x_i-\mu)^2}$$

  - Al logaritmo:

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

  - Derivando:

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

  - Otteniamo

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

La più probabile classificazione delle nuove istanza

- **MAP and MLE: Limitazioni**
  - **Il problema: "trovare l'ipotesi più probabile"**
  - **Ci basta la migliore classificazione di *x*, dato *D***
- **Metodi**
  - **Troviamo la migliore (MAP) *h*, la usiamo per classificare**
  - ***Non è detto che la classificazione sia ottimale***
  - ***Esempio:l'albero con l'errore minimo può sbagliare la classificazione su un'istanza specifica***
- **Obiettivo (raffinato)**
  - **Vogliamo determinare la più probabile classificazione**
  - **Combiniamo la predizione di tutte le ipotesi**
  - **Le predizioni sono pesati dalle probabilità condizionali**
  - **Result: Bayes Optimal Classifier**

Classificazione Bayesiana

- **Struttura**
  - Si trovi la più probabile classificazione
  - f: $X \rightarrow V$ (dominio ≡ spazio delle istanze, codominio ≡ insieme finito di valori)
  - $x \in X$ espresso come collezione di attributi $x \equiv (x_1, x_2, \ldots, x_n)$
  - classificatore Bayesiano
    - Dato x_i
    - Si determini: il più probabile valore $v_j \in V$

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid x) = \arg\max_{v_j \in V} P(v_j \mid x_1, x_2, \ldots, x_n)$$
$$= \arg\max_{v_j \in V} P(x_1, x_2, \ldots, x_n \mid v_j) P(v_j)$$

- **Problematiche**
  - Stimare $P(v_j)$ è semplice: per ogni valore v_j, contiamo la sua frequenza in $D = \{<x, f(x)>\}$
  - Tuttavia, è problematica la stima di $P(x_1, x_2, \ldots, x_n \mid v_j)$ senza assunzioni a priori

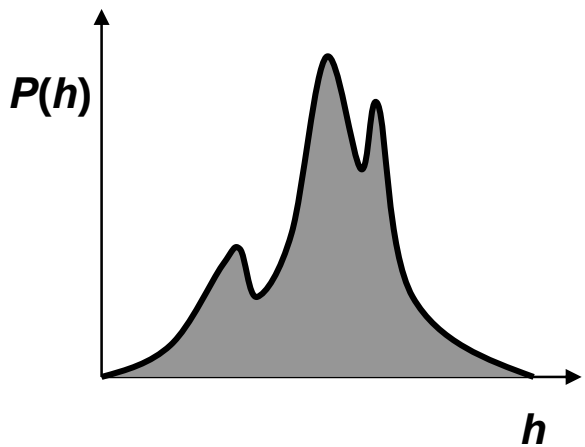
Classificazione Bayesiana (con ipotesi)

- **Idea**
  - $h_{MAP}(x)$ non fornisce necessariamente la classificazione più probabile
  - Esempio
    - Tre ipotesi: $P(h_1 \mid D) = 0.4$, $P(h_2 \mid D) = 0.3$, $P(h_3 \mid D) = 0.3$
    - Su una nuova istanza x, $h_1(x) = +$, $h_2(x) = -$, $h_3(x) = -$
    - Qual'è la migliore classificazione per x?
- **Bayes Optimal Classification (BOC)**

$$v^* = v_{BOC} = \arg\max_{v_j \in V} \sum_{h_i \in H} \left[P(v_j \mid h_i) \cdot P(h_i \mid D) \right]$$

  - Esempio
    - $P(h_1 \mid D) = 0.4$, $P(- \mid h_1) = 0$, $P(+ \mid h_1) = 1$
    - $P(h_2 \mid D) = 0.3$, $P(- \mid h_2) = 1$, $P(+ \mid h_2) = 0$
    - $P(h_3 \mid D) = 0.3$, $P(- \mid h_3) = 1$, $P(+ \mid h_3) = 0$
  - $\sum_{h_i \in H} \left[P(+ \mid h_i) \cdot P(h_i \mid D) \right] = 0.4$

    $\sum_{h_i \in H} \left[P(- \mid h_i) \cdot P(h_i \mid D) \right] = 0.6$
  - Risultato: $v^* = v_{BOC} = \arg\max_{v_j \in V} \sum_{h_i \in H} \left[P(v_j \mid h_i) \cdot P(h_i \mid D) \right] = -$



Naïve Bayes Classifier

- **Classificatore MAP** $v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid x) = \arg\max_{v_j \in V} P(v_j \mid x_1, x_2, \ldots, x_n)$
  $= \arg\max_{v_j \in V} P(x_1, x_2, \ldots, x_n \mid v_j) P(v_j)$
- **Naive Bayes**
  - Uno dei metodi più pratici di apprendimento
  - Assunzione di base: gli attributi di *x sono indipendenti*
- **Quando si può usare**
  - Il training set è grande
  - Gli attributi che descrivono *x* sono (sostanzialmente) independenti
- **Applicazione più di successo**
  - Classificazione di testi
- **Assunzione Naïve** $P(x_1, x_2, \ldots, x_n \mid v_j) = \prod_i P(x_i \mid v_j)$
- **Classificatore (Naïve) Bayes** $v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(x_i \mid v_j)$

Probabilità per PlayTennis

Outlook			Temperature			Humidity			Wind			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	**2**	**3**	**Hot**	**2**	**2**	**High**	**3**	**4**	**Light**	**6**	**2**	**9**	**5**
Overcast	**4**	**0**	**Mild**	**4**	**2**	**Normal**	**6**	**1**	**Strong**	**3**	**3**		
Rainy	**3**	**2**	**Cool**	**3**	**1**								
Sunny	**2/9**	**3/5**	**Hot**	**2/9**	**2/5**	**High**	**3/9**	**4/5**	**Light**	**6/9**	**2/5**	**9/14**	**5/14**
Overcast	**4/9**	**0/5**	**Mild**	**4/9**	**2/5**	**Normal**	**6/9**	**1/5**	**Strong**	**3/9**	**3/5**		
Rainy	**3/9**	**2/5**	**Cool**	**3/9**	**1/5**								

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Light	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Light	Yes
Rainy	Mild	High	Light	Yes
Rainy	Cool	Normal	Light	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Light	No
Sunny	Cool	Normal	Light	Yes
Rainy	Mild	Normal	Light	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Light	Yes
Rainy	Mild	High	Strong	No

Probabilità per PlayTennis

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	Light	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	Strong	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	Light	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	Strong	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

- **Una nuova istanza:**

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	Strong	?

verosimiglianza delle due classi

Per “yes” = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

Per “no” = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206

Probabilità associata:

P(“yes”) = 0.0053 / (0.0053 + 0.0206) = 0.205

P(“no”) = 0.0206 / (0.0053 + 0.0206) = 0.795

In pratica…

Outlook	Temp.	Humidity	Wind	Play
Sunny	Cool	High	Strong	?

← ***Evidenza E***

$$\Pr(yes \mid E) = \Pr(Outlook = Sunny \mid yes)$$
$$\times \Pr(Temperature = Cool \mid yes)$$
$$\times \Pr(Humidity = High \mid yes)$$
$$\times \Pr(Windy = True \mid yes)$$
$$\times \frac{\Pr(yes)}{\Pr(E)}$$

Probabilità della classe "yes"

$$= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{\Pr(E)}$$

Il problema della “frequenza-zero”

- **Che succede se il valore di un attributo non compare con un valore di classe?**
  (esempio: “Humidity = high” per la classe “yes”)
  - **La probabilità è zero!**

$$\Pr(Humidity = High \mid yes) = 0$$

  - **La probabilità a posteriori diventa zero!**

$$\Pr(yes \mid E) = 0$$

- **Rimedio: sommiamo 1 al conteggio di ogni combinazione attributo-classe (Laplace estimator)**

Stime di probabilità

- **Aggiungere una costante differente da 1 può risultare più appropriato**
- **Esempio su OutLook**

$$\frac{2+\mu/3}{9+\mu} \qquad \frac{4+\mu/3}{9+\mu} \qquad \frac{3+\mu/3}{9+\mu}$$

Sunny ***Overcast*** ***Rainy***

- **I pesi non devono necessariamente essere uguali (ma la somma deve essere 1)**

$$\frac{2+\mu p_1}{9+\mu} \qquad \frac{4+\mu p_2}{9+\mu} \qquad \frac{3+\mu p_3}{9+\mu}$$

Valori mancanti

- Nel training: l'istanza non viene conteggiata nella combinazione attributo-valore
- Nella classificazione: l'attributo viene omesso dal calcolo
- esempio:

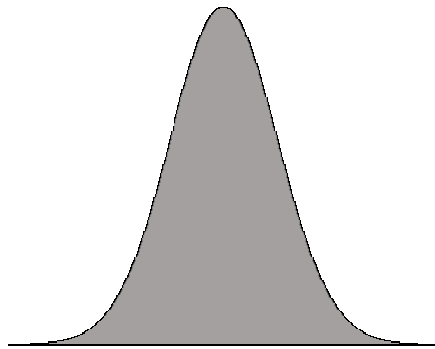
Outlook	Temp.	Humidity	Wind	Play
?	Cool	High	Strong	?

verosimiglianza di "yes" = 3/9 × 3/9 × 3/9 × 9/14 = 0.0238
verosimiglianza di "no" = 1/5 × 4/5 × 3/5 × 5/14 = 0.0343
P("yes") = 0.0238 / (0.0238 + 0.0343) = 41%
P("no") = 0.0343 / (0.0238 + 0.0343) = 59%

Attributi numerici

- **Assunzione tipica: gli attributi hanno una distribuzione normale (*Gaussiana)***
- **La densità di probabilità è definita da due parametri**
  - ***Valor medio sul campione*** μ

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$



  - ***Devianza sul campione*** σ

$$\sigma = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2$$

  - **densità *f(x):***

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Ancora Playtennis

Outlook			Temperature		Humidity		Wind			Play	
	Yes	*No*	*Yes*	*No*	*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	**2**	**3**	**64, 68,**	**65, 71,**	**65, 70,**	**70, 85,**	**Light**	**6**	**2**	**9**	**5**
Overcast	**4**	**0**	**69, 70,**	**72, 80,**	**70, 75,**	**90, 91,**	**Strong**	**3**	**3**		
Rainy	**3**	**2**	**72, …**	**85, …**	**80, …**	**95, …**					
Sunny	**2/9**	**3/5**	μ **=73**	μ **=75**	μ **=79**	μ **=86**	**Light**	**6/9**	**2/5**	**9/14**	**5/14**
Overcast	**4/9**	**0/5**	σ **=6.2**	σ **=7.9**	σ **=10.2**	σ **=9.7**	**Strong**	**3/9**	**3/5**		
Rainy	**3/9**	**2/5**									

- **Valore di densità**

$$f(temperature = 66 \mid yes) = \frac{1}{\sqrt{2\pi}\,6.2} e^{-\frac{(66-73)^2}{2*6.2^2}} = 0.0340$$

Classificazione su valori numerici

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Verosimiglianza di “yes” = 2/9 × 0.0340 × 0.0221 × 3/9 × 9/14 = 0.000036
Verosimiglianza di “no” = 3/5 × 0.0291 × 0.0380 × 3/5 × 5/14 = 0.000136
P(“yes”) = 0.000036 / (0.000036 + 0. 000136) = 20.9%
P(“no”) = 0.000136 / (0.000036 + 0. 000136) = 79.1%

Riassumendo…

- Il Naïve Bayes funziona sorprendentemente bene (anche sel'assunzione di indipendenza è chiaramente violata)
  - Non servono stime accurate di probabilità finché la probabilità massima è assegnata alla classe corretta
- Tuttavia: troppi attributi ridondanti può causare problemi
- Inoltre: molti attributi numerici non seguono la distribuzione normale
- Migliaramenti:
  - Selezioniamo i migliori attributi
  - Reti Bayesiane