## Lezione 6: Modelli lineari – classificazione

In questa lezione consideriamo modelli di classificazione nella forma

$$p(y=c/\vec{x};\vec{\vartheta}) = \frac{p(\vec{x}/y=c;\vec{\vartheta})\,p(y=c;\vec{\vartheta})}{\sum_{c'} p(\vec{x}/y=c';\vec{\vartheta})\,p(y=c';\vec{\vartheta})}$$

Il termine $p(y=c;\vec{\vartheta})$ è il prior sulle class labels ed il termine $p(\vec{x}/y=c;\vec{\vartheta})$ è chiamato CLASS CONDITIONAL DENSITY per la classe c. Il modello è chiamato "GENERATIVE CLASSIFIER" perché specifica un modo per generare le features $\vec{x}$ $\forall$ classe c campionando $p(\vec{x}/y=c;\vec{\vartheta})$. Invece, un "DISCRIMINATIVE CLASSIFIER" modella direttamente la classe posteriore $p(y/\vec{x};\vec{\vartheta})$. Scegliendo le probabilità condizionate in una certa maniera, il posteriore è una funzione lineare in $\vec{x}$ ($\ln p(y=c/\vec{x};\vec{\vartheta}) = w^T\vec{x} + \text{const.}$). Il metodo è chiamato LINEAR DISCRIMINANT ANALYSIS (LDA).

### 6.1 GAUSSIAN DISCRIMINANT ANALYSIS (GDA)

Consideriamo che la distribuzione di probabilità condizionata sulle classe sia una gaussiana

$$p(\vec{x}/y=c;\vec{\vartheta}) = \mathcal{N}(\vec{x}/\vec{\mu}_c, \Sigma_c)$$

Il corrispondente posteriore è pure gaussiano

$$p(\vec{y}=c/\vec{x};\vec{\vartheta}) \propto \pi_c \mathcal{N}(\vec{x}/\vec{\mu}_c, \Sigma_c)$$

dove $\pi_c = p(y=c)$ è la prior probability per la classe c. Prendendo il ln:

$$\ln p(\vec{y}=c/\vec{x};\vec{\vartheta}) = \ln \pi_c - \frac{1}{2}\ln|2\pi\Sigma_c| - \frac{1}{2}(\vec{x}-\vec{\mu}_c)^T\Sigma_c^{-1}(\vec{x}-\vec{\mu}_c) + \text{costante}$$

Questa funzione è chiamata FUNZIONE DISCRIMINANTE. Il DECISION BOUNDARY tra due classi (p. es. c e c') è una funzione quadratica di $\vec{x}$. Questo modello è noto come QUADRATIC DISCRIMINANT ANALYSIS (QDA).

Consideriamo ora un caso specifico di GDA in cui le matrici di covarianza siano condivise tra diverse classi, cioè $\Sigma_c = \Sigma$. Se $\Sigma$

è indipendente da $c$, riscriviamo:

$$\ln p(y=c/\vec{x};\vec{\vartheta}) = \ln \pi_c - \frac{1}{2}(\vec{x}-\vec{\mu}_c)^T \Sigma^{-1}(\vec{x}-\vec{\mu}_c) + \text{const}$$

$$= \underbrace{\ln \pi_c - \frac{1}{2}\vec{\mu}_c^T \Sigma^{-1}\vec{\mu}_c}_{\gamma_c} + \vec{x}^T \underbrace{\Sigma^{-1}\vec{\mu}_c}_{\beta_c} + \underbrace{\text{const} - \frac{1}{2}\vec{x}^T\Sigma^{-1}\vec{x}}_{K}$$

$$= \gamma_c + \vec{x}^T\beta_c + K$$ LINEAR DISCRIMINANT ANALYSIS (LDA)

Vediamo ora come fittare un modello GDA usando la MLE. La likelihood ha la forma

$$p(\mathcal{D}/\vec{\vartheta}) = \prod_{n=1}^{N_\mathcal{D}} \text{Cat}(y_n/\vec{\pi}) \prod_{c=1}^{C} \mathcal{N}(\vec{x}_n/\vec{\mu}_c, \Sigma_c)^{\mathbb{1}(y_n=c)}$$

$$\ln p(\mathcal{D}/\vec{\vartheta}) = \left[\sum_{n=1}^{N_\mathcal{D}}\sum_{c=1}^{C} \mathbb{1}(y_n=c)\ln \pi_c\right] + \sum_{c=1}^{C}\left[\sum_{n:y_n=c} \ln \mathcal{N}(\vec{x}_n/\vec{\mu}_c, \Sigma_c)\right]$$

pertanto è possibile ottimizzare separatamente $\pi$ e $(\mu_c, \Sigma_c)$. Sappiamo dalle lezioni precedenti che

$$\hat{\mu}_c = \frac{1}{N_{\mathcal{D}_c}}\sum_{n:y_n=c}\vec{x}_n \qquad \hat{\Sigma}_c = \frac{1}{N_{\mathcal{D}_c}}\sum_{n:y_n=c}(\vec{x}_n-\hat{\vec{\mu}}_c)(\vec{x}_n-\hat{\vec{\mu}}_c)^T$$

<u>Osservazioni</u>:

a) Se $\Sigma_c = \Sigma$, segue che $\hat{\Sigma} = \frac{1}{N_\mathcal{D}}\sum_{c=1}^{C}\sum_{n:y_n=c}(\vec{x}_n-\hat{\vec{\mu}}_c)(\vec{x}_n-\hat{\vec{\mu}}_c)^T$

b) Se forziamo $\Sigma_c$ ad essere diagonale, riduciamo il numero di parametri da $O(CD^2)$ a $O(CD)$. Questo è chiamato naive Bayes model. Quest'assunzione è alquanto restrittiva, ma funziona bene per $C, D$ grandi

c) Se assumiamo un prior uniforme sulle classi, possiamo calcolare la class label più probabile come segue

$$\hat{y}(\vec{x}) = \underset{c}{\text{argmax}} \ln p(y=c/\vec{x},\vec{\vartheta}) = \underset{c}{\text{argmin}}(\vec{x}-\vec{\mu}_c)^T\Sigma^{-1}(\vec{x}-\vec{\mu}_c)$$

Questo metodo è chiamato NEAREST CENTROID CLASSIFIER poiché assegniamo $\vec{x}$ alla classe con il più vicino $\vec{\mu}_c$

# 6.2 NAIVE BAYES CLASSIFIERS (NBC)

Supponiamo che le features siano condizionalmente indipendenti date le class labels. L'assunzione corrisponde ad usare una probabilità condizionata della forma

$$p(\vec{x}|y=c,\vec{\vartheta}) = \prod_{d=1}^{D} p(x_d|y=c,\vec{\vartheta}_{dc})$$

dove $\vec{\vartheta}_{dc}$ sono i parametri per la distribuzione di probabilità relativa alla classe c ed alla feature d. Quindi:

$$p(y=c|\vec{x},\vec{\vartheta}) = \frac{p(y=c|\pi)\prod_{d=1}^{D} p(x_d|y=c,\vec{\vartheta}_{dc})}{\sum_{c'} p(y=c'|\pi)\prod_{d=1}^{D} p(x_d|y=c',\vec{\vartheta}_{dc'})}$$

dove $\pi_c$ è il prior per la classe c e $\vec{\vartheta}=(\pi,\{\vec{\vartheta}_{dc}\})$ sono i parametri. Dobbiamo specificare la forma delle varie distribuzioni di probabilità.

- Nel caso di features binarie, $x_d \in \{0,1\}$ possiamo usare la distribuzione di Bernoulli: $p(\vec{x}|y=c,\vec{\vartheta}) = \prod_{d=1}^{D} \mathrm{Ber}(x_d|\vartheta_{dc})$ con $\vartheta_{dc}$ la probabilità che $x_d=1$ nella classe c.
- Nel caso di features categoriche, $x_d \in \{1,\dots,K\}$, possiamo usare la distribuzione categorica $p(\vec{x}|y=c,\vec{\vartheta}) = \prod_{d=1}^{D} \mathrm{Cat}(x_d|\vec{\vartheta}_{dck})$ con $\vartheta_{dck}$ la probabilità che $x_d=k$ dato $y=c$.
- Nel caso di features a valori reali, $x_d \in \mathbb{R}$, possiamo usare la distribuzione gaussiana $p(\vec{x}|y=c,\vec{\vartheta}) = \prod_{d=1}^{D} \mathcal{N}(x_d|\mu_{dc},\sigma^2_{dc})$ con $\mu_{dc}$ la media della feature d quando la class label è c e $\sigma^2_{dc}$ è la sua ~~var~~ varianza (questo caso è equivalente alla GDA con covarianze diagonali)

Fittiamo ora un NBC usando la MLE

$$p(\mathcal{D}|\vec{\vartheta}) = \prod_{n=1}^{N_D} \mathrm{Cat}(y_n|\pi)\prod_{d=1}^{D} p(x_{nd}|y_n,\vec{\vartheta}_d)$$

$$= \prod_{n=1}^{N_D} \mathrm{Cat}(y_n|\pi)\prod_{d=1}^{D}\prod_{c=1}^{C} p(x_{nd}|\vec{\vartheta}_{dc})^{\mathbb{I}(y_n=c)}$$

$$\ln p(\vec{\mathcal{D}}|\vec{\vartheta}) = \left[\sum_{n=1}^{N_D}\sum_{c=1}^{C}\mathbb{I}(y_n=c)\ln\pi_c\right] + \sum_{c=1}^{C}\sum_{d=1}^{D}\left[\sum_{n:y_n=c}\ln p(x_{nd}|\vec{\vartheta}_{dc})\right]$$

Questa equazione si decompone in due contributi

$$\ln p(\mathcal{D}/\vec{\vartheta}) = \ln p(\mathcal{D}_y/\pi) + \sum_c \sum_d \ln p(\mathcal{D}_{cd}/\vec{\vartheta}_{dc})$$

$\mathcal{D}_y = \{y_n : n = 1, N\}$ LABELS e $\mathcal{D}_{dc} = \{x_{nd} : y_n = c\}$ VALORI delle FEATURE d per samples di classe c

Nelle lezioni precedenti abbiamo mostrato che la MLE per $\pi$ è il vettore $\hat{\pi}_c = N_c/N$. Gli altri parametri sono stimati analogamente

- features binarie : $\hat{\vartheta}_{dc} = N_{dc}/N_c$
- features discrete ; $\hat{\vartheta}_{dck} = N_{dck}/N_c$
- features a valore reale;

$$\hat{\mu}_{dc} = \frac{1}{N_c} \sum_{n:y_n=c} x_{nd} \qquad \hat{\sigma}^2_{dc} = \frac{1}{N_c} \sum_{n:y_n=c} (x_{nd} - \hat{\mu}_{dc})^2$$

## 6.3. GENERATIVE VS DISCRIMINATIVE CLASSIFIERS

Abbiamo detto che un modello nella forma $p(\vec{x}, y) = p(y)p(\vec{x}/y)$ è detto GENERATIVE CLASSIFIER, perché genera esempi $\vec{x}$ $\forall$ classe $y$.
Un modello nella forma $p(y/\vec{x})$ è detto DISCRIMINATIVE CLASSIFIER perché può essere usato solamente per discriminare tra classi.

(A) Vantaggi dei DISCRIMINATIVE CLASSIFIERS

- Migliore accuratezza predittiva. La ragione è che stimare $p(y/\vec{x})$ è più semplice che stimare la probabilità congiunta $p(y, \vec{x})$.
- Libertà nel manipolare le features. I dati di input possono essere manipolati a piacere (e.g. basis expansion).
- Probabilità ben calibrate. Alcuni generative classifier (come NBC) fanno assunzioni particolarmente forti che possono essere sorgente di bias.

(B) Vantaggi dei GENERATIVE CLASSIFIERS

- Facili da fittare. MLE consiste nel contare e mediare.
- Possono facilmente fare a meno di features mancanti. (*)

- Possono fittare classi diverse separatamente
- Possono fare a meno del fatto che tutti i dati di training abbiano una label.
- Sono robusti rispetto a features spurie.

(*) Questo si può fare marginalizzando sulle features non note.

Ricordiamo che la distribuzione categorica è definita come

$$Cat(y/\vec{\vartheta}) \triangleq \prod_{c=1}^{C} \vartheta_c^{\mathbb{1}(y=c)}$$

cioè $p(y=c/\vec{\vartheta}) = \vartheta_c$. I parametri sono tali che $0 \leq \vartheta_c \leq 1$ e $\sum_{c=1}^{C} \vartheta_c = 1$

Ricordiamo che la distribuzione di Bernoulli è definita come

$$Ber(y/\vec{\vartheta}) = \begin{cases} 1-\vartheta & \text{se } y=0 \\ \vartheta & \text{se } y=1 \end{cases} \qquad Ber(y/\vec{\vartheta}) \triangleq \vartheta^y (1-\vartheta)^{1-y}$$