Algorytm oczekiwanie-maksymalizacja

Systemy uczące się - laboratorium

Mateusz Lango

Zakład Inteligentnych Systemów Wspomagania Decyzji
Wydział Informatyki i Telekomunikacji
Politechnika Poznańska

„Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”,
projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

Z ostatnich zajęć...

Problem

Zdefiniuj zadanie grupowania.

Problem

Jakie są typy algorytmów analizy skupień?

Problem

Jak działa algorytm k-średnich?

Problem

Jakie są wady algorytmu k-średnich?

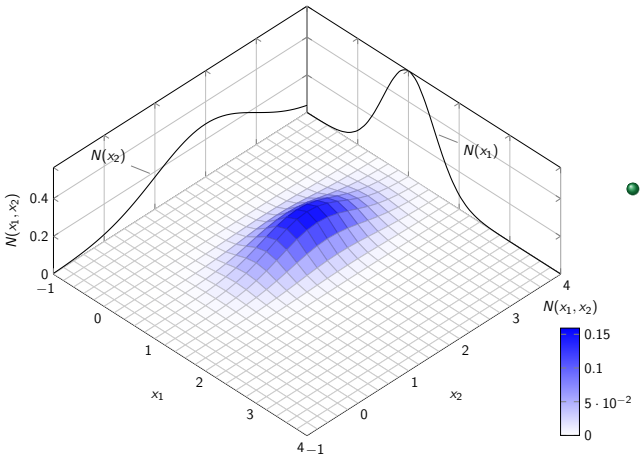
Klasyfikatory generatywne

- Na jednych z pierwszych zajęć wprowadziliśmy różnicę pomiędzy klasyfikatorami dyskryminacyjnymi i generatywnymi.

$$\max \sum_{i=1}^{n} \ln P(x_i, y_i) \qquad \text{vs.} \qquad \max \sum_{i=1}^{n} \ln P(y_i | x_i)$$

- Jedną z zalet algorytmów generatywnych jest m.in. możliwość stosunkowo łatwego(*) radzenia sobie z brakującymi danymi
- Czy jednak można skorzystać z klasyfikatorów generatywnych kiedy *brakuje wszystkich etykiet klas*?

Wielowymiarowy rozkład normalny



- Funkcja gęstości 1-wymiarowego rozkładu normalnego

$$N_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

- Funkcja gęstości d-wymiarowego rozkładu normalnego

$$N_{\vec{\mu},\Sigma}(\vec{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^\top \Sigma^{-1}(\vec{x}-\vec{\mu})\right)$$

parametryzowana wektorem średnich $\vec{\mu}$ oraz macierzą wariancji-kowariancji Σ

Powtórka: Liniowa analiza dyskryminacyjne

- Rozkład łączny $P(\vec{x}, y)$ możemy rozbić na dwie składowe korzystając z reguły łańcuchowej

$$P(\vec{x}, y) = P(\vec{x}|y)P(y)$$

- Zakładając rozkład normalny cech pod warunkiem klasy:

$$P(\vec{x}, y) = N(\vec{x}|\vec{\mu}_y, \Sigma_y)P(y)$$

$$N_{\vec{\mu},\Sigma}(\vec{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^\top \Sigma^{-1}(\vec{x} - \vec{\mu})\right)$$

- Dla każdej klasy potrzebujemy d średnich (dla każdej cechy) oraz kowariancje między parami cech (rzędu d^2)

Liniowa analiza dyskryminacyjna - estymacja

- Zgodnie z założeniami LDA:

$$P(\vec{x}, y) = N(\vec{x}|\vec{\mu}_y, \Sigma)P(y)$$

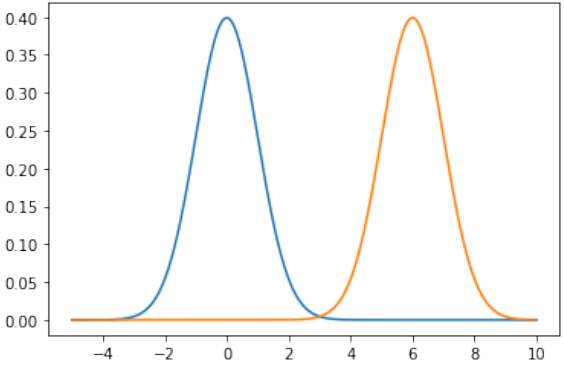
- Zgodnie z zasadą maksymalnej wiarygodności:

$$\max \sum_{i=1}^{n} \ln P(\vec{x_i}, y_i) = \sum_{i=1}^{n} \ln N(\vec{x}|\vec{\mu}_y, \Sigma)P(y)$$

- Przyrównując pochodną do 0 otrzymujemy:
  - $P(y) = \frac{n_y}{n}$ – liczba przykładów z klasy y podzielić przez liczbę wszystkich przykładów
  - $\mu_k = \frac{1}{n_k} \sum_{y_i=k} x_i$ dla każdej klasy k
  - $\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{y_i})(x_i - \mu_{y_i})^T$ (tylko 1, współdzielony między klasami) [1]

[1] Zwykle korzystamy z nieobciążonych estymatorów $\Sigma = \frac{1}{n-|C|} \sum_{k=1}^{|C|} \sum_{y_i=k} (x_i - \mu_{y_i})(x_i - \mu_{y_i})^T$ gdzie $|C|$ oznacza liczbę klas

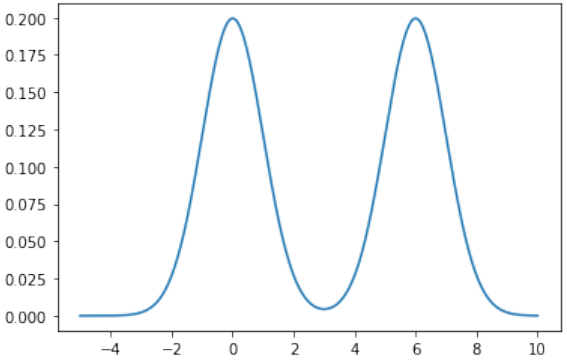
Liniowa analiza dyskryminacyjna - przypadek klasyfikacji binarnej



$$P(\vec{x}, y) = N(\vec{x}|\vec{\mu}_y, \Sigma)P(y)$$

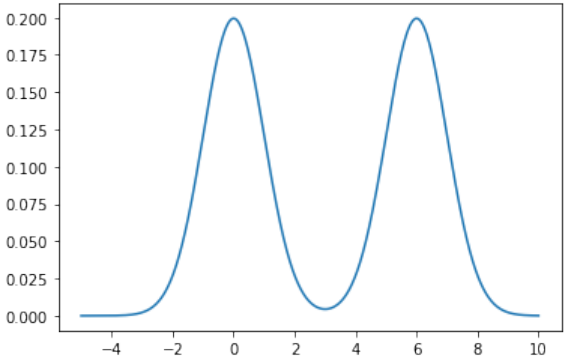
$$P(x) = ?$$

Liniowa analiza dyskryminacyjna - przypadek klasyfikacji binarnej



$$
\begin{aligned}
P(\vec{x}, y) &= N(\vec{x}|\vec{\mu}_y, \Sigma_y)P(y) \\
P(x) &= N(\vec{x}|\vec{\mu}_1, \Sigma_1)P(y=1) \\
&\quad + N(\vec{x}|\vec{\mu}_0, \Sigma_0)P(y=0) \\
P(y=1) &= ?
\end{aligned}
\tag{1}
$$

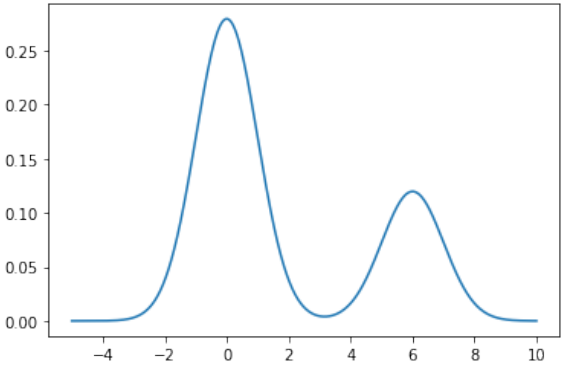
Liniowa analiza dyskryminacyjna - przypadek klasyfikacji binarnej



$$
\begin{aligned}
P(x) =& \pi N(\vec{x}|\vec{\mu}_1 = 0, \sigma_1 = 1) \\
&+ (1-\pi) N(\vec{x}|\vec{\mu}_0 = 6, \sigma_0 = 1)
\end{aligned}
\tag{2}
$$

$$\pi = P(y = 1) = 0.5$$

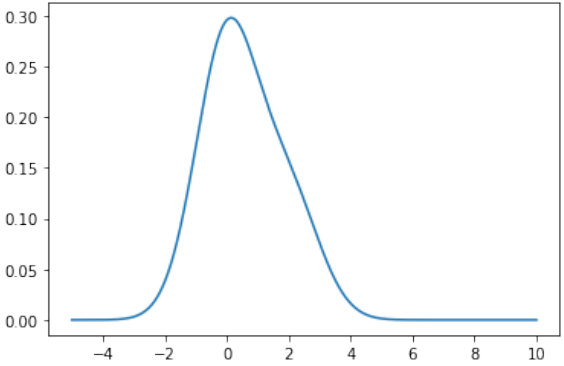
Liniowa analiza dyskryminacyjna - przypadek klasyfikacji binarnej



$$
\begin{aligned}
P(x) =& \pi N(x|\mu_1 = 0, \sigma_1 = 1) \\
&+ (1 - \pi) N(x|\mu_0 = 6, \sigma_0 = 1)
\end{aligned}
\tag{3}
$$

$$\pi = P(y = 1) = 0.7$$

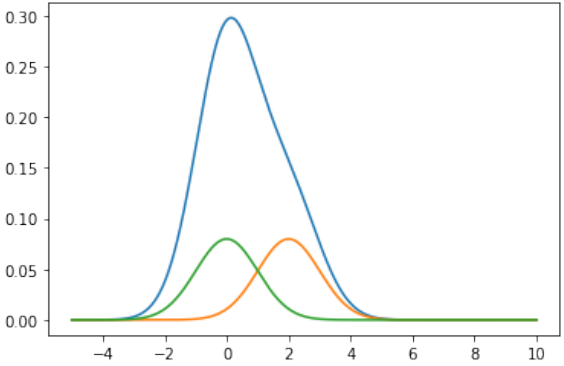
Liniowa analiza dyskryminacyjna - przypadek klasyfikacji binarnej



$$
\begin{aligned}
P(x) =& \pi N(x|\mu_1 = 0, \sigma_1 = 1) \\
&+ (1 - \pi) N(x|\mu_0 = 2, \sigma_0 = 1)
\end{aligned}
\tag{4}
$$

$$\pi = P(y = 1) = 0.7$$

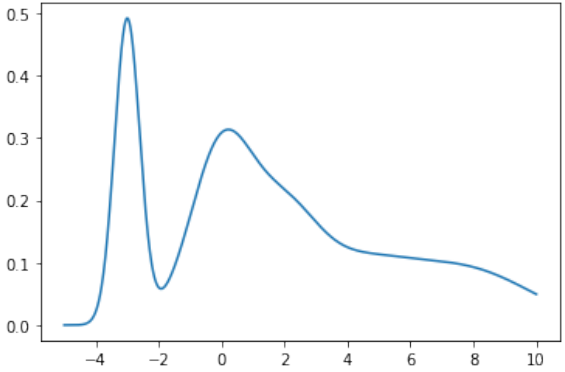
Liniowa analiza dyskryminacyjna - przypadek klasyfikacji binarnej



$$
\begin{aligned}
P(x) =& \pi N(x|\mu_1 = 0, \sigma_1 = 1) \\
&+ (1 - \pi) N(x|\mu_0 = 2, \sigma_0 = 1)
\end{aligned}
\tag{5}
$$

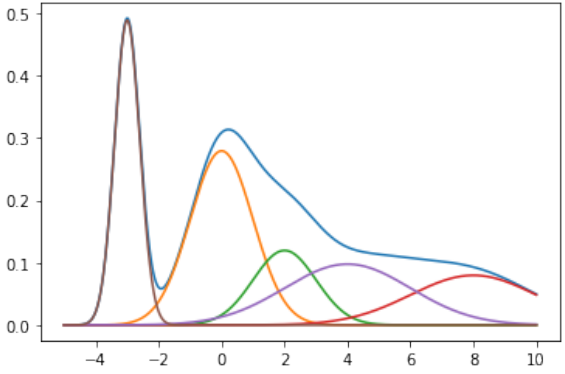
$$\pi = P(y = 1) = 0.7$$

Liniowa analiza dyskryminacyjna - przypadek klasyfikacji wieloklasowej



$$
\begin{aligned}
P(x) = & \pi_4 N(x|\mu_4 = -3, \Sigma_4 = 0.4) \\
& + ... \\
& + \pi_1 N(x|\mu_1 = 0, \sigma_1 = 1) \\
& + \pi_0 N(x|\mu_0 = 2, \sigma_0 = 1)
\end{aligned} \tag{6}
$$

Liniowa analiza dyskryminacyjna - przypadek klasyfikacji wieloklasowej



$$
\begin{aligned}
P(x) = & \pi_4 N(x|\mu_4 = -3\Sigma_4 = 0.4) \\
& + ... \\
& + \pi_1 N(x|\mu_1 = 0, \sigma_1 = 1) \\
& + \pi_0 N(x|\mu_0 = 2, \sigma_0 = 1)
\end{aligned} \tag{7}
$$

Mieszaniny rozkładów

$$P(x) = \sum_{i=1}^{K} \pi_i P_i(x)$$

przy założeniu $\sum_{i=1}^{K} \pi_i = 1$, $\pi_i \geq 0$

- statystyczna paralela do modeli addytywnych
- zwyczajowo P_i są rozkładami z tej samej rodziny
- „Gaussian mixture model is a *universal approximator of densities*, in the sense that any smooth density can be approximated with any specific nonzero amount of error by a Gaussian mixture model with enough components."
- problem identyfikowalności (ang. *identifiability*) (co najmniej $K!$ rozwiązań)

Mieszaniny rozkładów jako model statystycznej analizy skupień

$$P(x) = \sum_{i=1}^{K} \pi_i P_i(x)$$

przy założeniu $\sum_{i=1}^{K} \pi_i = 1$, $\pi_i \geq 0$

- zakładamy, że każdy komponent mieszaniny to jedno skupienie (grupa danych)
- jak uzyskać przypisanie do grupy (komponentu) dla obiektu?

$$P(z = 1|x) = \frac{\pi_1 P_i(x)}{\sum_{i=1}^{K} \pi_i P_i(x)}$$

Mieszaniny rozkładów jako model statystycznej analizy skupień

$$P(x) = \sum_{i=1}^{K} \pi_i P_i(x)$$

przy założeniu $\sum_{i=1}^{K} \pi_i = 1$, $\pi_i \geq 0$

- zakładamy, że każdy komponent mieszaniny to jedno skupienie (grupa danych)
- jak uzyskać przypisanie do grupy (komponentu) dla obiektu?

$$P(z = 1|x) = \frac{\pi_1 P_i(x)}{\sum_{i=1}^{K} \pi_i P_i(x)}$$

Jeśli $P_i(x)$ to rozkład normalny, a $K = 2$...

$$P(z = 1|x) = \frac{\pi N(x; \mu_1, \Sigma_1)}{(1 - \pi) N(x; \mu_0, \Sigma_0) + \pi N(x; \mu_1, \Sigma_1)}$$

Mieszaniny rozkładów jako model statystycznej analizy skupień

$$P(x) = \sum_{i=1}^{K} \pi_i P_i(x)$$

przy założeniu $\sum_{i=1}^{K} \pi_i = 1$, $\pi_i \geq 0$

- zakładamy, że każdy komponent mieszaniny to jedno skupienie (grupa danych)
- jak uzyskać przypisanie do grupy (komponentu) dla obiektu?

$$P(z = 1|x) = \frac{\pi_1 P_i(x)}{\sum_{i=1}^{K} \pi_i P_i(x)}$$

Jeśli $P_i(x)$ to rozkład normalny, a $K = 2$...

$$P(z = 1|x) = \frac{P(z = 1)N(x; \mu_1, \Sigma_1)}{(1 - P(z = 1))N(x; \mu_0, \Sigma_0) + P(z = 1)N(x; \mu_1, \Sigma_1)}$$

Mieszaniny rozkładów jako model statystycznej analizy skupień

$$P(x) = \sum_{i=1}^{K} \pi_i P_i(x)$$

przy założeniu $\sum_{i=1}^{K} \pi_i = 1$, $\pi_i \geq 0$

- zakładamy, że każdy komponent mieszaniny to jedno skupienie (grupa danych)
- jak uzyskać przypisanie do grupy (komponentu) dla obiektu?

$$P(z = 1|x) = \frac{\pi_1 P_i(x)}{\sum_{i=1}^{K} \pi_i P_i(x)}$$

Jeśli $P_i(x)$ zakłada niezależność warunkową, a $K = 2$...

$$P(z = 1|x) = \frac{\pi \prod_{j=1}^{d} P(x_j|z = 1)}{(1 - \pi) \prod_{j=1}^{d} P(x_j|z = 0) + \pi \prod_{j=1}^{d} P(x_j|z = 1)}$$

Mieszaniny rozkładów jako model statystycznej analizy skupień

$$P(x) = \sum_{i=1}^{K} \pi_i P_i(x)$$

przy założeniu $\sum_{i=1}^{K} \pi_i = 1$, $\pi_i \geq 0$

- zakładamy, że każdy komponent mieszaniny to jedno skupienie (grupa danych)
- jak uzyskać przypisanie do grupy (komponentu) dla obiektu?

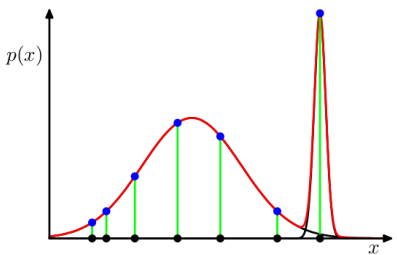
$$P(z = 1|x) = \frac{\pi_1 P_i(x)}{\sum_{i=1}^{K} \pi_i P_i(x)}$$

- przypisanie do grup jest probabilistyczne (tak jak klasyfikacja) tj. dany przykład może należeć do różnych komponentów (grup) z różnymi prawdopodobieństwami

Mieszanina rozkładów - wystarczy wysetymować?

$$\max \sum_{i=1}^{n} P(x_i) = \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_k P_k(x_i)$$

- LDA/QDA ma bardzo proste rozwiązanie, więc tutaj na pewno też?
- problem składający się gaussów



- brak „closed-form solution”
- brak unikalnego maksimum globalnego

Algorytm oczekiwanie-maksymalizacja

$$\max_{\theta} \ell(\theta) = \sum_{j=1}^{n} \log P(\vec{x}^{(j)}; \theta) = \sum_{j=1}^{n} \log \left[\sum_{y \in Y} P(\vec{x}^{(j)}, y; \theta) \right]$$

Algorytm oczekiwanie-maksymalizacja pomaga nam w optymalizacji wiarygodności zmarginalizowanego rozkładu $P(\vec{x}; \theta)$ o ile potrafimy efektywnie obliczyć dwa rozkłady: $P(y|\vec{x}; \theta)$ oraz $P(\vec{x}, y; \theta)$.

1. Algorytm rozpoczyna od pewnych (losowych) parametrów θ
2. Zastępujemy zmienne ukryte y ich wartościami oczekiwanymi ($P(y|\vec{x}; \theta)$)
3. Mając wartości oczekiwane y możemy wstawić je do $P(x, y; \theta)$ i zmaksymalizować
4. Skocz do kroku 2

Algorytm oczekiwanie-maksymalizacja

1. Zainicjalizuj $t = 0$ oraz parametry modelu $\theta^{(t)}$
2. Dopóki nie jest spełniony warunek stopu:
   1. Wyznacz WARTOŚĆ OCZEKIWANĄ logarytmu funkcji wiarygodności po zmiennych ukrytych przy znajomości $P(Y|\vec{X};\theta^{(t)})$.

$$\mathbb{E}_{y\sim P(Y|\vec{X};\theta^{(t)})} \log P(\vec{x}, y; \theta)$$

   2. Zaktualizuj θ poprzez MAKSYMALIZACJĘ wcześniej wyznaczonej funkcji oraz zinkrementuj t

$$\theta^{(t+1)} = \arg\max_{\theta} \mathbb{E}_{y\sim P(y|\vec{x};\theta^{(t)})} \log P(\vec{x}, y; \theta)$$

Co tak naprawdę oznacza krok E?

$$\mathbb{E}_{y\sim P(y|\vec{x};\theta^{(t)})} \log P(\vec{x}, y; \theta) = \sum_{y\in Y} P(y|\vec{x};\theta^{(t)}) \log P(\vec{x}, y; \theta)$$

Algorytm oczekiwanie-maksymalizacja

1. Zainicjalizuj $t = 0$ oraz parametry modelu $\theta^{(t)}$
2. Dopóki nie jest spełniony warunek stopu:
   1. Wyznacz WARTOŚĆ OCZEKIWANĄ logarytmu funkcji wiarygodności po zmiennych ukrytych przy znajomości $P(Y|\vec{X};\theta^{(t)})$.

$$\mathbb{E}_{y\sim P(Y|\vec{X};\theta^{(t)})} \log P(\vec{x}, y; \theta)$$

   2. Zaktualizuj θ poprzez MAKSYMALIZACJĘ wcześniej wyznaczonej funkcji oraz zinkrementuj t

$$\theta^{(t+1)} = \arg\max_{\theta} \mathbb{E}_{y\sim P(y|\vec{x};\theta^{(t)})} \log P(\vec{x}, y; \theta)$$

Co tak naprawdę oznacza krok E?

$$\mathbb{E}_{y\sim P(y|\vec{x};\theta^{(t)})} \log P(\vec{x}, y; \theta) = \sum_{y\in Y} P(y|\vec{x};\theta^{(t)}) \log P(\vec{x}, y; \theta)$$

EM dla GMM

- Krok E

$$\widetilde{z_i} = \mathbb{E}[z_i] = P(z_i = 1|x_i) = \frac{\pi N(x_i; \mu_1, \Sigma_1)}{(1-\pi)N(x_i; \mu_0, \Sigma_0) + \pi N(x_i; \mu_1, \Sigma_1)}$$

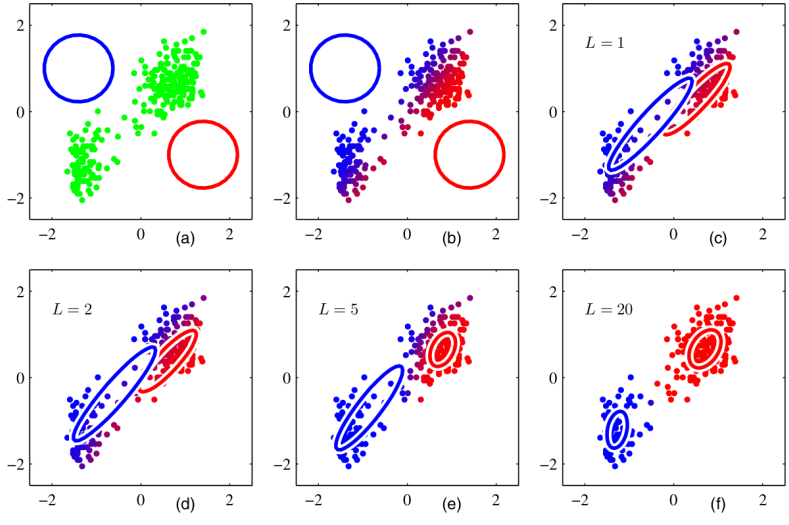
- Krok M

$$\pi = \frac{\sum_{i=1}^{n} z_i}{n}$$

$$\mu_1 = \frac{\sum_{i=1}^{n} z_i x_i}{\sum_{i=1}^{n} z_i}$$

$$\ldots$$

Przykład



EM-GMM w praktyce

- duże podobieństwo do k-średnich (przypadek graniczny $\Sigma = I\sigma^2$ oraz $\sigma \to 0$)
- Jak wybrać liczbę komponentów K?
- Jak poradzić sobie ze składającymi się gaussami?
- Jak zainicjalizować?
- Ile parametrów ma model?

Zadanie

Problem

Dany jest zestaw danych w którym każda obserwacja zawiera wyniki 10 rzutów monetą, w których część wyników została wygenerowana przez podrzucanie monety A, a druga część wyników poprzez podrzucanie monety B. Na koniec eksperymentu okazało się, że monety nie były dobrze wyważone – zaprojektuj algorytm EM, który podzieli obserwacje na te wygenerowane monetą A i B.

Problem

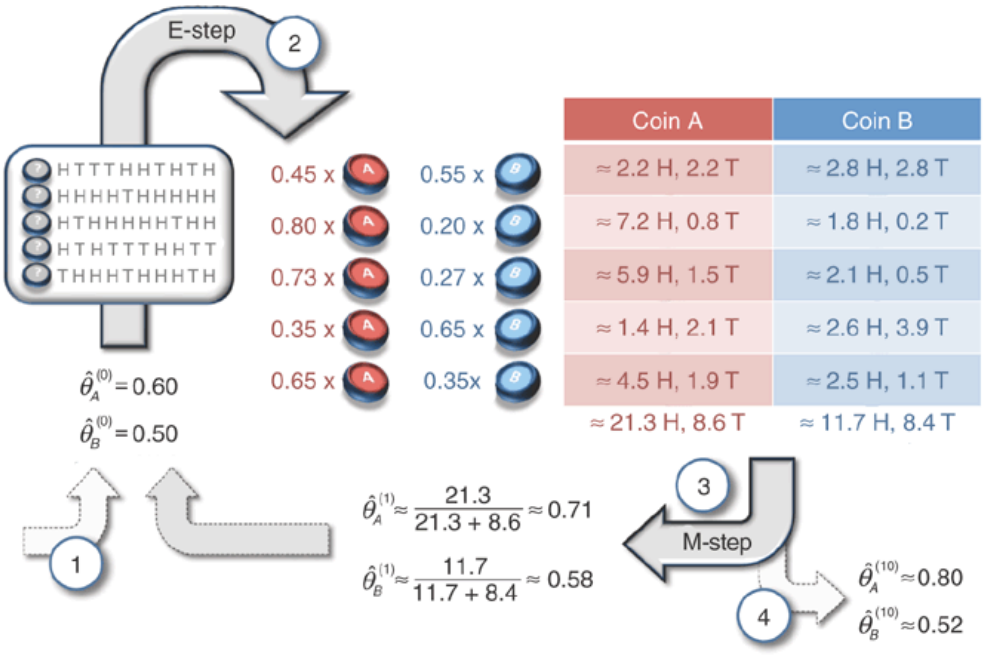
Dany jest klasyfikator Naiwnego Bayesa dla cech binarnych i klasyfikacji binarnej. Zaprojektuj algorytm EM uczący „klasyfikator" na danych nienadzorowanych.

a Maximum likelihood

	Coin A	Coin B
B H T T T H H T H T H		5 H, 5 T
A H H H H T H H H H H	9 H, 1 T	
A H T H H H H H T H H	8 H, 2 T	
B H T H T T T H H T T		4 H, 6 T
A T H H H T H H H T H	7 H, 3 T	
	24 H, 6 T	9 H, 11 T

$$\hat{\theta}_A = \frac{24}{24+6} = 0.80$$

$$\hat{\theta}_B = \frac{9}{9+11} = 0.45$$


E-step
2
H T T T H H T H T H
H H H H T H H H H H
H T H H H H H T H H
H T H T T T H H T T
T H H H T H H H T H
0.45 x A
0.80 x A
0.73 x A
0.35 x A
0.65 x A
0.55 x B
0.20 x B
0.27 x B
0.65 x B
0.35x B
Coin A
Coin B
≈ 2.2 H, 2.2 T
≈ 2.8 H, 2.8 T
≈ 7.2 H, 0.8 T
≈ 1.8 H, 0.2 T
≈ 5.9 H, 1.5 T
≈ 2.1 H, 0.5 T
≈ 1.4 H, 2.1 T
≈ 2.6 H, 3.9 T
≈ 4.5 H, 1.9 T
≈ 2.5 H, 1.1 T
≈ 21.3 H, 8.6 T
≈ 11.7 H, 8.4 T
$\hat{\theta}_A^{(0)} = 0.60$
$\hat{\theta}_B^{(0)} = 0.50$
1
$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3 + 8.6} \approx 0.71$
$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7 + 8.4} \approx 0.58$
3
M-step
4
$\hat{\theta}_A^{(10)} \approx 0.80$
$\hat{\theta}_B^{(10)} \approx 0.52$

Uczenie pół-nadzorowane

- Dempster et. all 1977 $\Rightarrow$ Little 1977 (!)
- Teoria: z wystarczająco dużą liczbą przykładów niezaetykietowanych , znajdziemy EM bardziej prawdopodobny model niż używając tylko danych nadzorowanych
- Jeśli model jest prawdziwy (pasuje do procesu generującego dane) to bardziej prawdopodobny $\Rightarrow$ bardziej trafny (Bayes optimal)
- Jednak w praktyce...

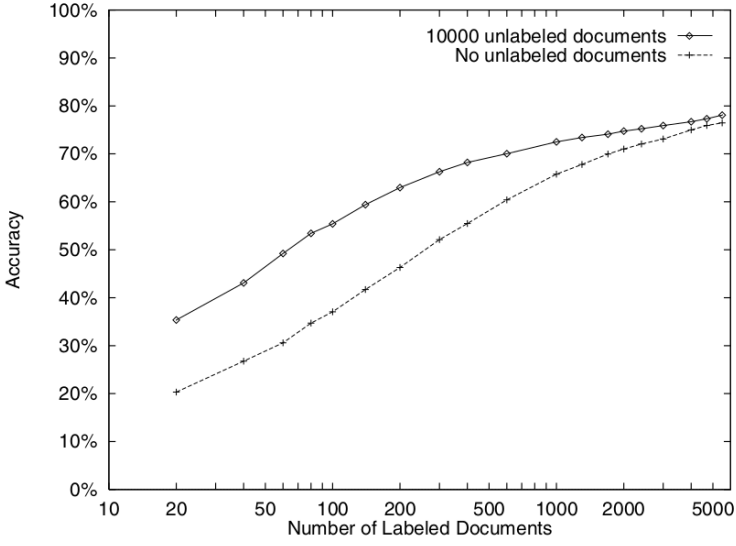
Pół-nadzorowany Naiwny Bayes

$$\max \sum_{i=1}^{n_L} \ln P(\vec{x}_i, y_i) + \sum_{i=1}^{n_U} \ln P(\vec{x}_i)$$

Algorithm 3.1 Basic EM algorithm for semi-supervised learning of a text classifier

- **Inputs:** Collections X_l of labeled documents and X_u of unlabeled documents.
- Build an initial naive Bayes classifier, $\hat{\theta}$, from the labeled documents, X_l, only. Use maximum a posteriori parameter estimation to find $\hat{\theta} = \arg\max_\theta \mathrm{P}(X_l|\theta)\mathrm{P}(\theta)$ (see Eqs. 3.5 and 3.6).
- Loop while classifier parameters improve, as measured by the change in $l(\theta|X, Y)$ (the log probability of the labeled and unlabeled data, and the prior) (see Equation 3.8):
  - **(E step)** Use the current classifier, $\hat{\theta}$, to estimate component membership of each unlabeled document, i.e., the probability that each mixture component (and class) generated each document, $\mathrm{P}(c_j|x_i; \hat{\theta})$ (see Eq. 3.7).
  - **(M step)** Re-estimate the classifier, $\hat{\theta}$, given the estimated component membership of each document. Use maximum a posteriori parameter estimation to find $\hat{\theta} = \arg\max_\theta \mathrm{P}(X, Y|\theta)\mathrm{P}(\theta)$ (see Eqs. 3.5 and 3.6).
- **Output:** A classifier, $\hat{\theta}$, that takes an unlabeled document and predicts a class label.

Pół-nadzorowany Naiwny Bayes



Pół-nadzorowany Naiwny Bayes

Category	NB1	EM1	NB*	EM*
acq	86.9	81.3	88.0 (4)	**93.1** (10)
corn	94.6	93.2	96.0 (10)	**97.2** (40)
crude	94.3	94.9	95.7 (13)	**96.3** (10)
earn	94.9	95.2	**95.9** (5)	95.7 (10)
grain	94.1	93.6	96.2 (3)	**96.9** (20)
interest	91.8	87.6	95.3 (5)	**95.8** (10)
money-fx	93.0	90.4	94.1 (5)	**95.0** (15)
ship	94.9	94.1	**96.3** (3)	95.9 (3)
trade	91.8	90.2	94.3 (5)	**95.0** (20)
wheat	94.0	94.5	96.2 (4)	**97.8** (40)

Dziękuję za uwagę!

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego