1. Теоретичні відомості

1.1. Мультиномінальна модель та максимальна апостеріорна ймовірності

В мультиномінальній моделі, на відміну від моделі Бернулі, враховується кількість входжень кожного слова в документ. $w_k, k = 1, .., D$–всі унікальні слова в корпусі. // x^k–кількість повторень слова w_k в документі θ_y^k-ймовірність зустрічі слова w_k на позиції

$$\widehat{y}(x) = argmax_y P(y|x) = argmax_y P(y)P(x|y) = (1)$$

$$= argmax_y P(y) \frac{(\sum_{k=1}^{D} x^k)!}{\sqcap_{k=1}^{D} x^k!} \sqcap_{k=1}^{D} (\theta_y^k)^{x^k} = (2)$$

$$= argmax_y lnP(y) + \sum_{k=1}^{D} (x^k ln\theta_y^k)$$

де $P(y) = \frac{N_y}{N}, \theta_y^k = \frac{n_{yk}+\alpha}{n_y+\alpha D}$–емпіричні оцінкі ймовірностей. N_y–кількість документів класу y n_y–кількість слів в документах класу y n_{yk}–кількість повторень слова w_k в документах класу y (1)–Байєсівське правило апостеріорної ймовірності (2)–припущення "наївно байєса"

1.2. Припущення про позиційну незалежність

Розглянемо множину об'єктів $D = d_1, .., d_m$, кожен з яких володія набором ознак з множини всіх ознак $F = f_1, .., f_q$. Модель наївного баїсівського класифікатора приймає два припущення:

- 1. порядок признаків об'єкта не має значення
- 2. ймовірності признаків не залежать один від одного при заданому классі: $P(f_i \cap f_j|c) = P(f_j|c)$.

1.3. Застосування для інформаційного пошуку

Модель байєсівського наївного классифікатору часто використовується для інформаційного пошуку та классифікації текстів, осткільки вона часто дає хороші результати, хоч породжені нею оцінкі часто далекі від ідеальних. Хоч зазвичай в моделі, яка називається наївною, ознаки розглядаються як категоріальні, або випадкові, величини з розподіли Бернулі, варіанти, в яких використовується мультиноміальний розподіл, як правило краще моделють входження слів в документ. Врахувати якісні ознаки можна двома способами: моделюючи їх як як нормально розподілені в межах кожного классу чи за допомогою параметричної оцінки щільності.