02-2.Байесовская статистика и автоматическая обработка текстов

Статистический вывод: языковая модель

- Много данных – нужно уметь использовать данные
- Языковые модели
- Обработка речи
- Предсказание следующего слова по имеющейся цепочке
- Формула Байеса, Байесовский вывод

Теория вероятностей

- Как вероятно некоторое событие
- Пространство событий Ω
- Событие A -подмножество Ω
- Вероятность (функция, распределение)

$$P:\Omega \to [0,1]$$

Вероятность полной совокупности событий

Прямые и обратные задачи

- Прямая задача: в урне лежат 10 шаров, из них 3 черных. Какова вероятность выбрать черный шар?

- Обратная задача: перед нами две урны, в каждой – по 10 шаров, но в одной 3 черных, а в другой – 6. Кто-то взял из какой-то урны шар, и он оказался черным. Какова вероятность, что он брал шар из первой урны?

- Обратные задачи содержат скрытые переменные (номер урны, из которой брали шар)

Пример: автоматическая рубрикация текстов

- Эксперты отнесли тексты к рубрикам рубрикатора
- Это сделано на основе каких-то слов в тексте
- Слов в текстах много
- Вопрос: к каким рубрикам относятся конкретные слова текста (и с какой вероятностью)

Априорная вероятность

- Вероятность до рассмотрения дополнительного знания

$$P(A)$$

Условная вероятность

- Иногда мы имеем частичное знание о событиях
- Условная или апостериорная вероятность
- Предположим, что мы знаем, что совершилось событие B
- Вероятность, что совершилось событие A при условии знания о событии B $P(A|B)$

Условная вероятность-2

$$P(A,B) = P(A \mid B)P(B)$$
$$= P(B \mid A)P(A)$$

- Вероятность совершения двух событий A и B

- Цепь условных вероятностей

P(A,B,C,D…) = P(A)P(B|A)P(C|A,B)P(D|A,B,C..)

- Два события A и B независимы, если

P(A) = P(A|B)

Теорема Байеса

- Теорема Байеса переворачивает зависимость между событиями
- Мы видели, что $P(A|B) = \dfrac{P(A,B)}{P(B)}$
- Теорема Байеса:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Теорема Байеса-2

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(B)=P(B|A)P(A)+P\ (B|\bar{A})P(\bar{A})$

$P(B)=\sum i\ (P(B|Ai)P(Ai))$

Пример

- S:несгибаемость шеи, M: менингит
- P(S|M) =0.5, P(M) = 1/50,000
  P(S)=1/20
- Если у человека не сгибается шея, то какая вероятность менингита?

Пример

- S:несгибаемость шеи, M: менингит
- P(S|M) =0.5, P(M) = 1/50,000
  P(S)=1/20
- Если у человека не сгибается шея, то какая вероятность менингита?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)}$$

$$= \frac{0.5 \times 1/50{,}000}{1/20} = 0.0002$$

Языковые модели

- В общем случае, для многих языковых явлений вероятность P неизвестна
- Нужно оценить P, (или модель M языка)

- Чтобы сделать такую оценку, нужно рассмотреть данные и на этом основании построить предположения о вероятности

Оценка P

- Частотный подход
- Байесовский подход

Вероятность как частота

- Обычно в классической теории вероятности вероятность понимается как предел отношения количества определенного результата эксперимента к общему количеству экспериментов

$$f_u = \frac{C(u)}{N}$$

- Приложения
  - Физика
  - Азартные игры

- Стандартный пример: бросание монетки
- Относительная частотность стабилизируется при бесконечном повторении результата у некоторого числа

Байесовская статистика

- Во многих случаях невозможно говорить о большом количестве экспериментов

- Байесовская статистика оценивает степень доверия
- Степени доверия вычисляются, начиная с априорных предположений (априорная вероятность, априорное вероятностное распределение)
- и затем корректируются на основе данных с использованием теоремы Байеса (апостериорное распределение)

Пример различия подходов

- Бросаем монетку (возможно, фальшивую) 10 раз
- Выпало 8 (i) орлов
- Какова вероятность выпадения орла в следующем бросании
- Частотный подход
  - P (орел) = i/(i+j) – 0.8
- Байесовский подход
  - P (орел)= (i+1)/ (i+j+2) – 0.75
  - Правило Лапласа
  - Вывод: http://logic.pdmi.ras.ru/~sergey/teaching/mlbayes/01-inference.pdf (Страница С.Николенко)

Восстановление модели

- Так называемые параметрические методы
- Модель
  - Предполагаем, что данные распределены по какому-то закону (распределение)
    - Равномерно
    - Экспоненциально Θ
    - Биномиально
  - Распределение имеет параметры

Математическое ожидание

$$p(x) = p(X = x) = p(A_x)$$

$$A_x = \{\omega \in \Omega : X(\omega) = x\}$$

$$\sum_x p(x) = 1 \qquad 0 \le p(x) \le 1$$

- Математическое ожидание – это среднее значение случайной величины

$$E(x) = \sum_x x p(x) = \mu$$

Дисперсия

- Дисперсия случайной величины – это мера разброса случайной величины, т.е. отклонения от математического ожидания
- σ – стандартное отклонение

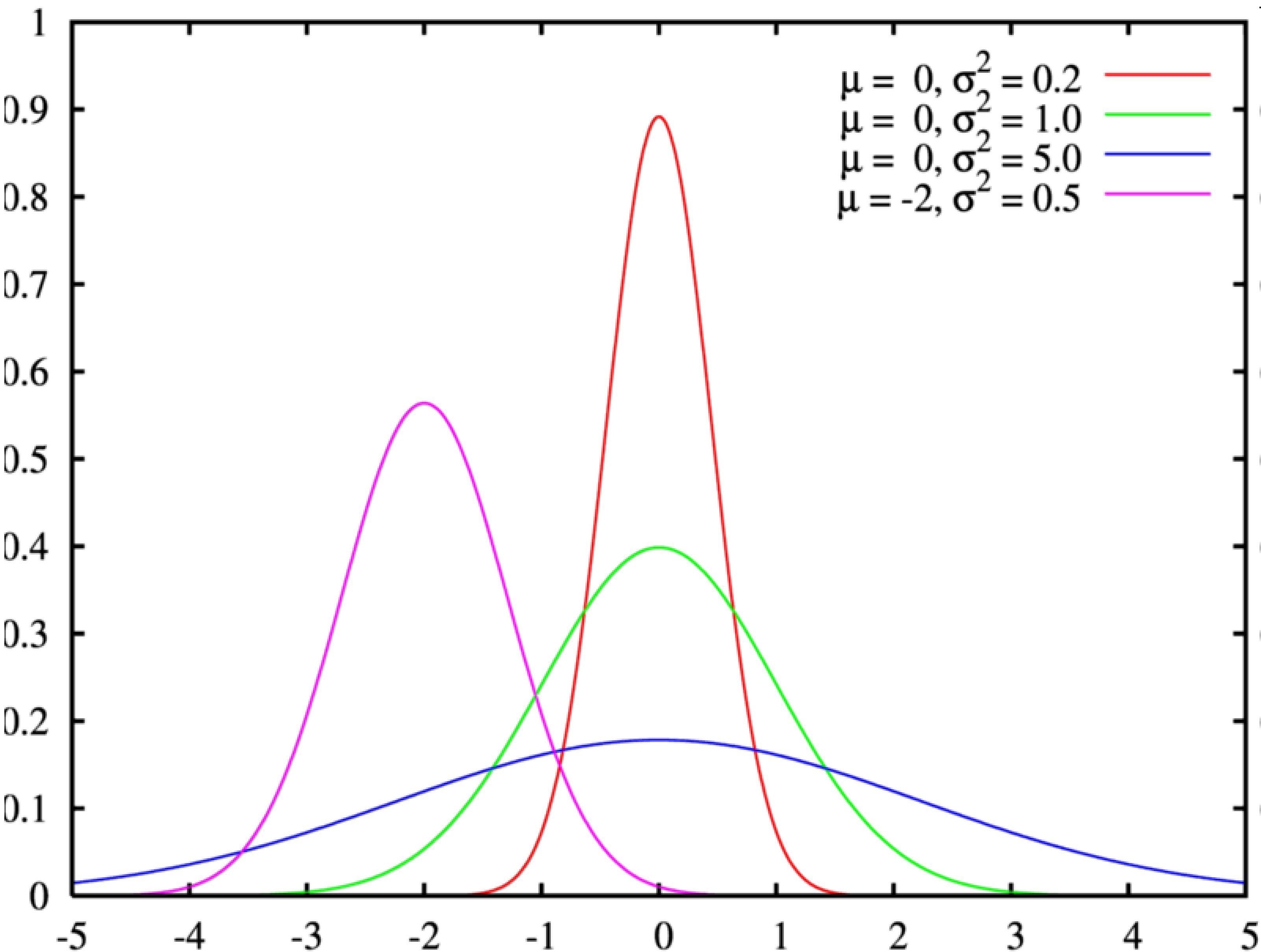
$$Var(X) = E((X - E(X))^2)$$
$$= E(X^2) - E^2(X) = \sigma^2$$

Нормальное распределение

- Непрерывное
- Два параметра: математическое ожидание μ и стандартное отклонение σ

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

Нормальное распределение



Частотный подход

- Выбор модели – выбирается на основе сравнения максимального правдоподобия $\overset{*}{M}$

$$\overset{*}{M} = \underset{M}{\operatorname{argmax}} P\left(D \mid M, \overset{*}{\theta}(M) \right)$$

$$\overset{*}{\theta} = \underset{\theta}{\operatorname{argmax}} P(D \mid M, \theta)$$

Байесовская статистика

$$\overset{*}{M} = \underset{M}{\operatorname{argmax}}\, P(M|D)$$

$$= \underset{M}{\operatorname{argmax}}\, \frac{P(D|M)P(M)}{P(D)}$$

$$= \underset{M}{\operatorname{argmax}}\, P(D|M)P(M)$$

MAP!

MAP is maximum a posteriori

Лингвистическая задача

- Лингвист Джон интересуется редкой лингвистической конструкцией, которая встречается в среднем один раз на 100 тысяч предложений.
- Джон придумал специальный шаблон, который находит такую конструкцию.
- Если в предложении есть эта конструкция, то шаблон обнаружит ее с вероятностью 0.95
- Если нет, то он может ошибочно показать ее с вероятностью 0.005.
- Тест говорит, что в предложении обнаружена эта конструкция. Какова вероятность, что – это правда?

Задачки

- Книгу пишут два автора. Иван написал 40% текста, а Петр 60%. В среднем на 9 страницах из тысячи Иван делает ошибку, А Петр – на 1 странице из 250. На одной случайно открытой странице обнаружилась ошибка. Какова вероятность, что ее допустил Иван.

- Три лингвиста делают морфологическую разметку предложения. Один лингвист сделал 40% всей работы, два другие – по 30%. Первый лингвист ошибается в 0.02% сложных случаев разметки, второй лингвист – 0.03%, третий – 0.01%. В выдаче встретился неправильный пример разметки. Какова вероятность, что ошибку сделал первый лингвист?