Behind LDA

Часть 1

Кольцов С.Н.

Различия в подходах к теории вероятностей

Случайная величина — это величина, которая принимает в результате опыта одно из множества значений, причём появление того или иного значения этой величины до её измерения нельзя точно предсказать.

1. В частотном подходе (классический подход) предполагается, что случайность есть объективная неопределенность . Вероятность рассчитывается из серии экспериментов и является мерой случайности как эмпирической данности. Исторически частотный подход возник из практической задачи: анализа азартных игр — области, в которой понятие серии испытаний имеет простой и ясный смысл.
2. В байесовском подходе предполагается, что случайность характеризует наше незнания. Например, случайность при бросании кости связана с незнанием динамических характеристик игральной кости, сопротивления воздуха и так далее. Многие задачи частотным методом решить невозможно (точнее, вероятность искомого события строго равна нулю). В то же время интерпретация вероятности как меры нашего незнания позволяет получить отличный от нуля осмысленный ответ.

Понятие вероятности

Вероятность события — Вероятностью события А называют отношение числа благоприятствующих этому событию исходов к общему числу всех равновозможных несовместных элементарных исходов.

Например. Вероятность того, что на кубике выпадет четное число, равна следующему отношению P=3/6=1/2.

Понятие условной вероятности

Условной вероятностью события **А** при условии, что произошло событие **В**, называется число **P(A|B)=P(B, A)/ P(B)**,
P(B, A) – произведение вероятностей, **P(B)** – вероятность события **В.**

Например. В урне 3 белых и 3 черных шара. Из урны дважды вынимают по одному шару, не возвращая их обратно. Найти вероятность появления белого шара при втором испытании (событие В), если при первом испытании был извлечен черный шар (событие А).
Вероятность события А=3/6=1/2
Произведение вероятностей P(B, A) =(3/6)*(3/5)=9/30

Итоговый результат: (9/30)/(1/2)=3/5

Формула Байеса

Байесовская вероятность — это интерпретация понятия вероятности, используемое в байесовской теории. Вероятность определяется как степень уверенности в истинности суждения**.**

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

P(A) — априорная вероятность гипотезы *A (заранее известная вероятность)*;
P (A|B) — вероятность гипотезы *A* при наступлении события *B* **(апостериорная вероятность**);
P(B|A) — вероятность наступления события *B* при истинности гипотезы *A*;
P(B) — полная вероятность наступления события *B*.
P(A|B)) — вероятность наступления события *A* при истинности гипотезы *B*;

Формула Байеса позволяет «переставить причину и следствие»: по известному факту события вычислить вероятность того, что оно было вызвано данной причиной. Таким образом, формула Байеса может быть использована для разработки алгоритмов классификации.

Априорные и апостериорные суждения

1. Предположим, мы хотим узнать значение некоторой неизвестной величины.
2. У нас имеются некоторые знания, полученные до (a priori) наблюдений/эксперимента. Это может быть опыт прошлых наблюдений, какие-то модельные гипотезы, ожидания.
3. В процессе наблюдений эти знания подвергаются постепенному уточнению. После (a posteriori) наблюдений/эксперимента у нас формируются новые знания о явлении.
4.Будем считать, что мы пытаемся оценить неизвестное значение величины P(A|B) посредством наблюдений некоторых ее косвенных характеристик (гипотез).

Формула Байеса (1763 г.) устанавливает правила, по которым происходит преобразование знаний в процессе наблюдений.

Пример применения формулы Байеса в E-Health

Пример: случайному пациенту сделали тест на наличие СПИД, и получили положительный результат. Пусть точность теста 99.8% (т.е. он дает положительный результат у 0.2% здоровых людей). Какова вероятность, что у этого пациента СПИД?

Априорная вероятность P(больной) – доля больных в стране (пусть 0.3%)

$$P(\text{больной} \mid \text{тест}+) = \frac{P(\text{тест}+ \mid \text{больной}) \cdot P(\text{больной})}{P(\text{тест}+ \mid \text{больной})P(\text{больной}) + P(\text{тест}+ \mid \text{здоровый})P(\text{здоровый})} =$$

$$= \frac{1 \cdot 0.003}{1 \cdot 0.003 + 1 \cdot 0.002} = 60\%$$

Вероятностная постановка задачи классификации

Пусть имеется множество объектов X и конечное множество классов Y . Требуется построить алгоритм способный классифицировать произвольный объект X в рамках заданного множества Y.
Апостериорная вероятность принадлежности объекта X классу Y по формуле Байеса:

$$P(X \mid Y) = \frac{p(X,Y)}{P(X)} = \frac{p(X)P(Y \mid X)}{P(X)}$$

$P(X \mid Y)$ - Апостериорная вероятность

$p(X,Y)$ - Априорная вероятность

Задача классификации заключается в расчете (оценке) апостериорной информации на основании априорной информации. Такая оценка может быть реализована при помощи формулы Байеса. Однако существует проблема оценивания априорной величины p(x,y)

Задача восстановления априорного распределения

p(x,y)

Оценка функции ***p(x,y)*** может быть реализован при помощи трех методов.

1. Непараметрическое восстановление плотности основано на локальной аппроксимации плотности p(x) в окрестности классифицируемого объекта x ∈ X. Пример, Алгоритм Парзена-Розенблатта (метод парзеновского окна).

2. Параметрическое восстановление плотности основано на предположении, что плотность распределения известна с точностью до параметра, ***p(x,y)*** = φ(x; θ), где φ фиксированная функция. Пример. Нормальный дискриминантный анализ. LSA – в основе лежит метод SVD разложения.

3. Восстановление смеси плотностей. Если функцию плотности ***p(x,y)*** не удаётся смоделировать параметрическим распределением, можно попытаться описать её смесью нескольких распределений:

Собственно именно третий метод является основой LDA

$$p(x) = \sum_{j=1}^{k} w_j \varphi(x; \theta_j), \quad \sum_{j=1}^{k} w_j = 1,$$

Latent Dirichlet allocation

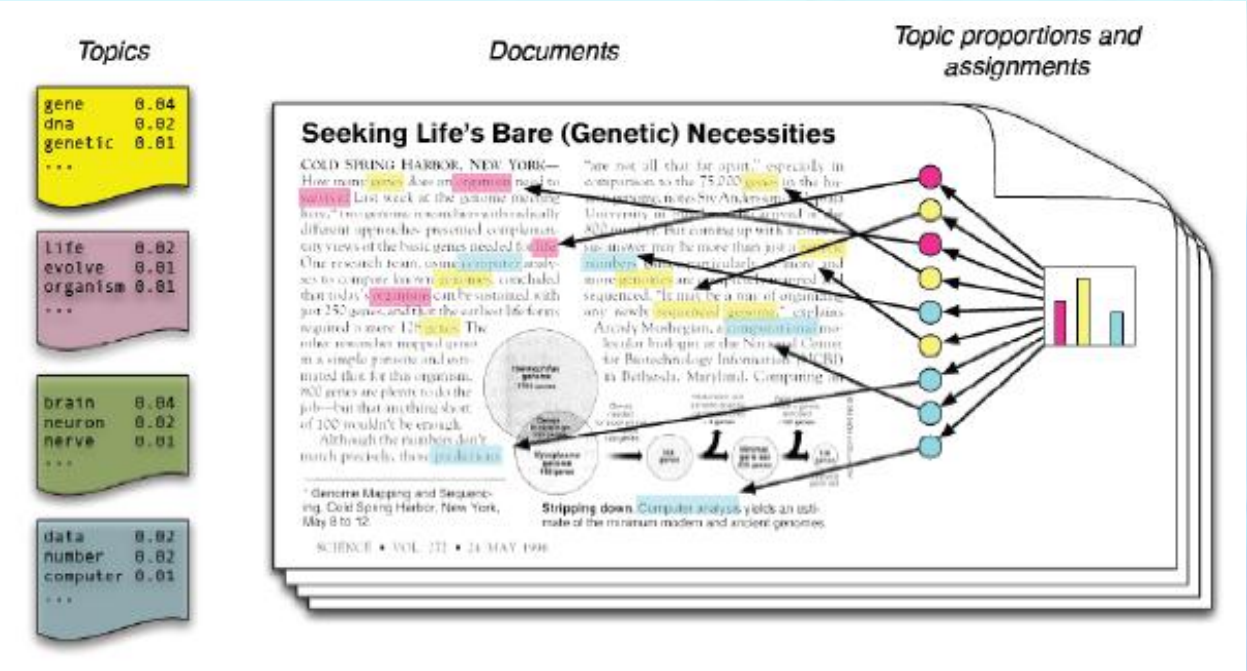
Основное предположение тематической модели Latent Dirichlet Allocation состоит в том, что каждый документ с некоторой вероятностью может принадлежать множеству тематик. Тема эта совокупность слов, где каждое слово имеет некоторую вероятность принадлежности к данной тематике.

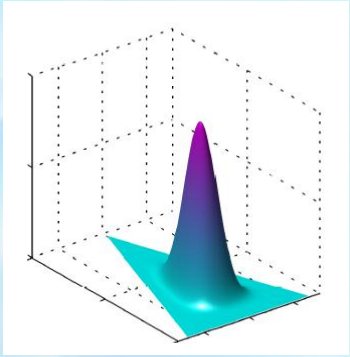
Формально тема определяется как дискретное (мультиномиальное) вероятностное распределение в пространстве слов заданного словаря.

Тематическим моделированием называется решение задачи, обратной классификации . Каждый документ в корпусе текстов рассматривается как наблюдаемая случайная независимая выборка слов (мешок слов), порождённая некоторым, скрытым (латентным) множеством тем. По этим данным требуется восстановить вероятностные распределения всех тем в корпусе и определить, каким именно подмножеством тем порождён каждый документ.

Тематическое моделирование основано на применении формулы Байеса, в которой распределение слов и тем выражено в виде смеси плотностей распределений слов и документов.

Модель LDA





$$p(W,Z,\Theta\,|\,\Phi,\alpha)=\prod_{d=1}^{D}p(\theta_d\,|\,\alpha)\prod_{n=1}^{N}p(w_{d,n}\,|\,z_{d,n},\Phi)p(z_{d,n}\,|\,\theta_d)$$

$p(z_{d,n}\,|\,\theta_d)=\Theta_{d,z_{d,n}}$ - функция распределение вероятности тематик в документах

$p(w_{d,n}\,|\,z_{d,n},\Phi)=\Phi_{z_{d,n},w_{d,n}}$ - функция распределения вероятности слов по темам и документам

Логика вычисления LDA (безотносительно к методу расчета)

$$p(z_{d,n} \mid \theta_d) = \Theta_{d,z_{d,n}}$$

$$p(w_{d,n} \mid z_{d,n}, \Phi) = \Phi_{z_{d,n},w_{d,n}}$$

$$p(W, Z, \Theta \mid \Phi, \alpha)$$

$$P(X \mid Y) = \frac{p(X,Y)}{P(X)}$$

1. Задаем начальное приближение функций распределения вероятности тематик в документах

2. Задаем начальное приближение функций распределения вероятности слов по темам и документам.

3. Рассчитываем вероятность вынимания слов из мешка.

4. Выдергиваем слова из реальных документов и сравниваем расчетные и экспериментальные вероятности.

5. Если разница между сгенерированной вероятностью и вероятностью выдернутого слова из текста больше наперед заданной величины, то производим коррекцию начальных распределений и переходим на шаг 1.

6. Если разница между сгенерированной вероятностью и вероятностью выдернутого слова из текста меньше наперед заданной величины, то наше начальное приближение хорошо описывает наши документы. Расчет закончен.