Лекция 1. Байесовские и марковские сети

Д. П. Ветров[1] Д. А. Кропотов[2] А. А. Осокин[1]

[1]МГУ, ВМиК, каф. ММП

[2]ВЦ РАН

Курс «Графические модели»

План

Программа курса

- Курс состоит из 10 лекций
- В рамках курса будет предложено 3 задания для самостоятельной работы
- В рамках курса будет 3 семинарских занятия с выдачей домашнего задания на каждом из них
- Будет одна контрольная работа по курсу
- Выполнение заданий и написание контрольной является необходимым условием для допуска к экзамену

Основной предмет изучения

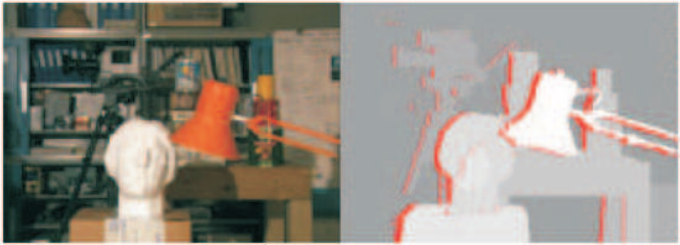
Основные обеъкты изучения

- Аппарат графических моделей (байесовские и марковские сети)
- Аппарат байесовского вывода
- Некоторые методы дискретной оптимизации
- Методы структурного обучения

Примеры применения графических моделей

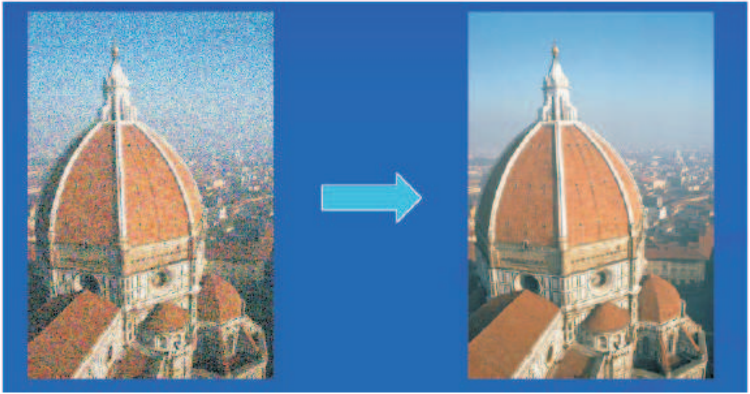
- Семантическая сегментация изображений
- Обработка видео
- Распознавание речи
- Анализ социальных сетей
- Трехмерная реконструкция
- Задачи трекинга
- Биоинформатика
- Органическая химия
- Декодирование зашумленных сообщений
- и многое другое

Примеры использования



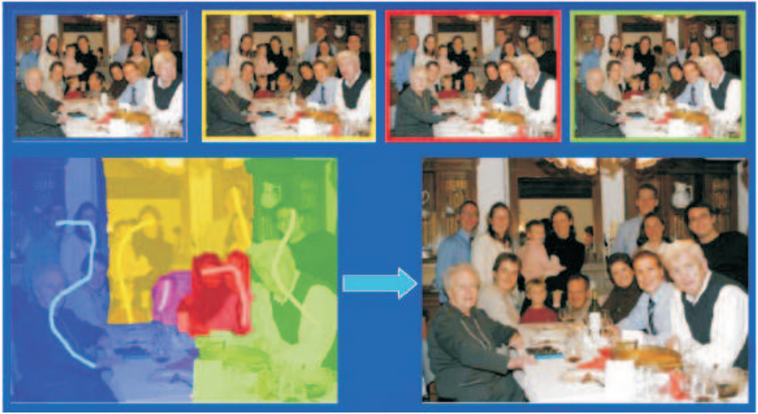
Оценка диспаритетов и восстановление стерео

Примеры использования



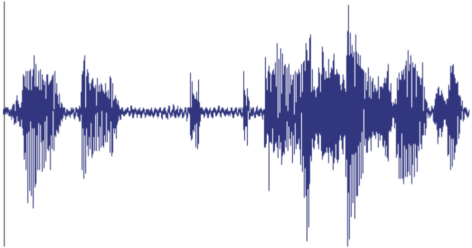
Фильтрация изображений

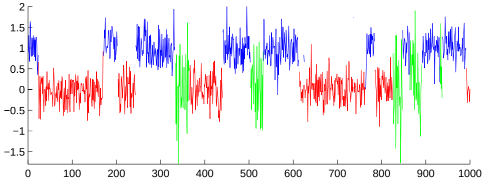
Примеры использования



Создание коллажей

Примеры использования





Обработка сигналов и распознавание речи

Примеры использования

u	LITTLE-JACK-HORNER-SAT-IN-THE-CORNER-EATING-A-CHRISTMAS-PIE--HE-PUT-IN-H
v	RIDE-A-COCK-HORSE-TO-BANBURY-CROSS-TO-SEE-A-FINE-LADY-UPON-A-WHITE-HORSE
matches:	.*....*..******.*..............*...........*................*...........

Восстановление зашумленных сообщений в ситуациях, когда уровень шума превышает возможности самокоррекции

Матричная нотация

- При работе с многомерными величинами очень удобна матричная нотация, т.е. представление многих операций над векторами и числами в виде операций над матрицами
- Скалярное произведение двух векторов $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ принимает вид

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{d} x_i y_i = \boldsymbol{x}^T \boldsymbol{y},$$

т.е. вектора трактуются как частные случаи матриц
- Квадратичная форма

$$\langle A\boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{d} \sum_{j=1}^{d} x_i a_{ij} y_j = \boldsymbol{x}^T A \boldsymbol{y}$$

- Матричная нотация облегчает математические выкладки и позволяет реализовать вычисления на ЭВМ более эффективно

Пример использования

- Предположим нам надо решить несовместную систему линейных уравнений $A\boldsymbol{x} \approx \boldsymbol{b}$, $A \in \mathbb{R}^{m \times n}$
- Для этого будем минимизировать квадрат нормы невязки (система-то нерешаемая) $\|A\boldsymbol{x} - \boldsymbol{b}\|^2 \to \min_{\boldsymbol{x}}$
- Представляя норму в матричной виде, дифференцируя по вектору и приравнивая производную к нулю получаем известную формулу для псевдорешения СЛАУ

$$\|A\boldsymbol{x} - \boldsymbol{b}\|^2 = \langle A\boldsymbol{x} - \boldsymbol{b}, A\boldsymbol{x} - \boldsymbol{b}\rangle = (A\boldsymbol{x} - \boldsymbol{b})^T(A\boldsymbol{x} - \boldsymbol{b}) =$$
$$(A\boldsymbol{x})^T A\boldsymbol{x} - \boldsymbol{b}^T A\boldsymbol{x} - (A\boldsymbol{x})^T\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{b} =$$
$$\boldsymbol{x}^T A^T A\boldsymbol{x} - 2\boldsymbol{x}^T A^T\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{b}$$
$$\frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{x}^T A^T A\boldsymbol{x} - 2\boldsymbol{x}^T A^T\boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{b}\right) = 2A^T A\boldsymbol{x} - 2A^T\boldsymbol{b} = 0$$
$$\boldsymbol{x} = \left(A^T A\right)^{-1} A^T\boldsymbol{b}$$

- Заметим, что если матрица A квадратная (число уравнений равно числу неизвестных) и невырожденная, то последняя формула переходит в формулу обычного решения СЛАУ $\boldsymbol{x} = A^{-1}\boldsymbol{b}$

Задача условной оптимизации

Пусть $f(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ — гладкая функция. Предположим, что нам необходимо найти ее экстремум:

$$f(\boldsymbol{x}) \to \operatorname*{extr}_{\boldsymbol{x}}$$

Для того, чтобы найти экстремум (решить задачу безусловной оптимизации), достаточно проверить условие стационарности:

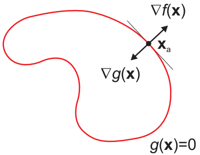
$$\nabla f(\boldsymbol{x}) = 0$$

Предположим, что нам необходимо найти экстремум функции при ограничениях:

$$f(\boldsymbol{x}) \to \operatorname*{extr}_{\boldsymbol{x}}$$

$$g(\boldsymbol{x}) = 0$$

Поверхность ограничения



Заметим, что $\nabla g(\boldsymbol{x})$ ортогонален поверхности ограничения $g(\boldsymbol{x}) = 0$. Пусть $\boldsymbol{x}$ и $\boldsymbol{x} + \boldsymbol{\varepsilon}$ — две близкие точки поверхности. Тогда

$$g(\boldsymbol{x} + \boldsymbol{\varepsilon}) \simeq g(\boldsymbol{x}) + \boldsymbol{\varepsilon}^T \nabla g(\boldsymbol{x})$$

Т.к. $g(\boldsymbol{x} + \boldsymbol{\varepsilon}) = g(\boldsymbol{x})$, то $\boldsymbol{\varepsilon}^T \nabla g(\boldsymbol{x}) \simeq 0$. При стремлении $\|\boldsymbol{\varepsilon}\| \to 0$ получаем $\boldsymbol{\varepsilon}^T \nabla g(\boldsymbol{x}) = 0$. Т.к. $\boldsymbol{\varepsilon}$ параллелен поверхности $g(\boldsymbol{x}) = 0$, то $\nabla g(\boldsymbol{x})$ является нормалью к этой поверхности.

Функция Лагранжа

Необходимым условием оптимальности является ортогональность $\nabla f(\boldsymbol{x})$ поверхности ограничения, т.е.:

$$\nabla f + \lambda \nabla g = 0 \tag{1}$$

Здесь $\lambda \neq 0$ — коэффициент Лагранжа. Он может быть любого знака.

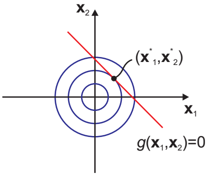
Функция Лагранжа

$$L(\boldsymbol{x}, \lambda) \triangleq f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$

Тогда

$$\begin{aligned} \nabla_{\boldsymbol{x}} L = 0 \quad & \Rightarrow \quad \text{условие (1)} \\ \frac{\partial}{\partial \lambda} L = 0 \quad & \Rightarrow \quad g(\boldsymbol{x}) = 0 \end{aligned}$$

Функция Лагранжа. Пример.



$$f(x_1,x_2) = 1 - x_1^2 - x_2^2 \to \max_{x_1,x_2}$$
$$g(x_1,x_2) = x_1 + x_2 - 1 = 0$$

Функция Лагранжа:

$$L(\boldsymbol{x},\lambda) = 1 - x_1^2 - x_2^2 + \lambda(x_1 + x_2 - 1)$$

Условия стационарности:

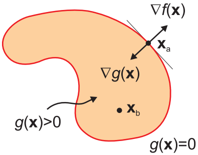
$$-2x_1 + \lambda = 0$$
$$-2x_2 + \lambda = 0$$
$$x_1 + x_2 - 1 = 0$$

Решение: $(x_1^*, x_2^*) = (\frac{1}{2}, \frac{1}{2}),\ \lambda = 1.$

Ограничение в виде неравенства

$$f(\boldsymbol{x}) \to \max_{\boldsymbol{x}}$$

$$g(\boldsymbol{x}) \geq 0$$



Решение	Ограничение	Условие стационарности
Внутри области $g(\boldsymbol{x}) > 0$	неактивно	$\nabla f(\boldsymbol{x}) = 0$, $\nabla_{\boldsymbol{x}} L = 0$, $\lambda = 0$
На границе $g(\boldsymbol{x}) = 0$	активно	$\nabla f(\boldsymbol{x}) = -\lambda \nabla g(\boldsymbol{x})$, $\nabla_{\boldsymbol{x},\lambda} L = 0$, $\lambda > 0$

Условие дополняющей нежесткости:

$$\lambda g(\boldsymbol{x}) = 0$$

Теорема Каруша-Куна-Таккера

Пусть $f_i : X \to \mathbb{R}$, $i = 0, 1, \ldots, m$ — выпуклые функции, отображающие нормированное пространство X в прямую, $A \in X$ — выпуклое множество. Рассмотрим следующую задачу оптимизации:

$$f_0(\boldsymbol{x}) \to \min; \quad f_i(\boldsymbol{x}) \leq 0, \ i = 1, \ldots, m, \ \boldsymbol{x} \in A \tag{2}$$

Теорема

1. Если $\hat{\boldsymbol{x}} \in \mathrm{absmin}\,(2)$ — решение задачи, то найдется ненулевой вектор множителей Лагранжа $\boldsymbol{\lambda} \in \mathbb{R}^{m+1}$ такой, что для функции Лагранжа $L(\boldsymbol{x}) = \sum_{i=0}^{m} \lambda_i f_i(\boldsymbol{x})$ выполняются условия:
   a) стационарности $\min_{\boldsymbol{x} \in A} L(\boldsymbol{x}) = L(\hat{\boldsymbol{x}})$
   b) дополняющей нежесткости $\lambda_i f_i(\hat{\boldsymbol{x}}) = 0, \ i = 1, \ldots, m$
   c) неотрицательности $\lambda_i \geq 0$
2. Если для допустимой точки $\hat{\boldsymbol{x}}$ выполняются условия a)–c) и $\lambda_0 \neq 0$, то $\hat{\boldsymbol{x}} \in \mathrm{absmin}\,(2)$
3. Если для допустимой точки $\hat{\boldsymbol{x}}$ выполняются условия a)–c) и $\exists \tilde{\boldsymbol{x}} \in A : f_i(\tilde{\boldsymbol{x}}) < 0, i = 0, \ldots, m$ (условие Слейтера), то $\hat{\boldsymbol{x}} \in \mathrm{absmin}\,(2)$

План

Краткое напоминание основных вероятностных понятий

- $X : \Omega \to \mathbb{R}$ – случайная величина
- Вероятность попадания величины в интервал (a, b) равна

$$P(a \le X \le b) = \int_a^b p(x)dx,$$

где $p(x)$ – плотность распределения X,

$$p(x) \ge 0, \quad \int_{-\infty}^{\infty} p(x)dx = 1$$

- Если поведение случайной величины определяется некоторым параметром, возникают условные плотности $p(x|\theta)$. Если рассматривать условную плотность как функцию от параметра

$$f(\theta) = p(x|\theta),$$

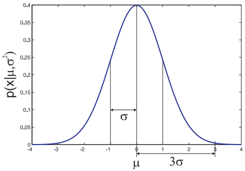
то принято говорить о т.н. функции правдоподобия

Нормальное распределение

- Нормальное распределение играет важнейшую роль в математической статистике

$$X \sim \mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\mu = \mathbb{E}X, \quad \sigma^2 = \mathbb{D}X \triangleq \mathbb{E}(X-\mathbb{E}X)^2$$



- Из центральной предельной теоремы следует, что сумма независимых случайных величин с ограниченной дисперсией стремится к нормальному распределению
- На практике многие случайные величины можно считать приближенно нормальными

Многомерное нормальное распределение

- Многомерное нормальное распределение имеет вид

$$X \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi}^n \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right),$$

  где $\boldsymbol{\mu} = \mathbb{E}X$, $\Sigma = \mathbb{E}(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^T$ — вектор математических ожиданий каждой из n компонент и матрица ковариаций соответственно
- Матрица ковариаций показывает, насколько сильно связаны (коррелируют) компоненты многомерного нормального распределения

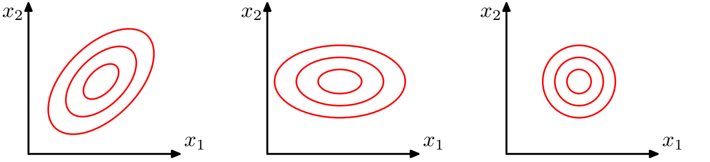
$$\Sigma_{ij} = \mathbb{E}(X_i - \mu_i)(X_j - \mu_j) = \mathrm{Cov}(X_i, X_j)$$

- Если мы поделим ковариацию на корень из произведений дисперсий, то получим коэффициент корреляции

$$\rho(X_i, X_j) \triangleq \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathbb{D}X_i \mathbb{D}X_j}} \in [-1, 1]$$

Особенности нормального распределения

- Нормальное распределение полностью задается первыми двумя моментами (мат. ожидание и матрица ковариаций/дисперсия)
- Матрица ковариаций неотрицательно определена, причем на диагоналях стоят дисперсии соответствующих компонент
- Нормальное распределение имеет очень легкие хвосты: большие отклонения от мат. ожидания практически невозможны. Это обстоятельство нужно учитывать при приближении произвольных случайных величин нормальными



(a) (b) (c)

Основная задача мат. статистики

- Распределение случайной величины X известно с точностью до параметра θ
- Имеется выборка значений величины X, $\boldsymbol{x} = (x_1, \dots, x_n)$
- Требуется оценить значение θ
- Метод максимального правдоподобия

$$\hat{\theta}_{ML} = \arg\max f(\theta) = \arg\max p(\boldsymbol{x}|\theta) = \arg\max \prod_{i=1}^{n} p(x_i|\theta)$$

- Можно показать, что ММП (не путать с одноименной кафедрой) является асимптотически оптимальным при $n \to \infty$
- Обычно максимизируют не само правдоподобие, а его логарифм, т.к. это вычислительно проще (произведение плотностей по всем объектам переходит в сумму логарифмов плотностей)

Пример использования

- Пусть имееется выборка из нормального распределения $\mathcal{N}(x|\mu,\sigma^2)$ с неизвестными мат. ожиданием и дисперсией
- Выписываем логарифм функции правдоподобия

$$L(X|\mu,\sigma) = -\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2} - n\log\sigma - \frac{n}{2}\log(2\pi) \to \max_{\mu,\sigma}$$

$$\frac{\partial L}{\partial\mu} = -\sum_{i=1}^{n}\frac{(x_i-\mu)}{\sigma^2} = 0 \qquad \frac{\partial L}{\partial\sigma} = \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^3} - \frac{n}{\sigma} = 0$$

$$\mu_{ML} = \frac{1}{n}\sum_{i=1}^{n}x_i \qquad \sigma^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^2$$

План

Условная вероятность

- Пусть X и Y — случайные величины с плотностями $p(x)$ и $p(y)$ соответственно
- В общем случае их совместная плотность $p(x, y) \neq p(x)p(y)$. Если это равенство выполняется, величины называют независимыми
- Условной плотностью называется величина

$$p(x|y) = \frac{p(x, y)}{p(y)}$$

- Смысл: как факт $Y = y$ влияет на распределение X. Заметим, что $\int p(x|y)dx \equiv 1$, но $\int p(x|y)dy$ не обязан равняться единице, т.к. относительно y это не плотность, а функция правдоподобия
- Очевидная система тождеств $p(x|y)p(y) = p(x, y) = p(y|x)p(x)$ позволяет легко переходить от $p(x|y)$ к $p(y|x)$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Правило суммирования вероятностей

- Все операции над вероятностями базируются на применении всего двух правил
- Правило суммирования: Пусть $A_1, \ldots, A_k$ взаимоисключающие события, одно из которых всегда происходит. Тогда

$$P(A_i \cup A_j) = P(A_i) + P(A_j) \qquad \sum_{i=1}^{k} P(A_i) = 1$$

- Очевидное следствие (формула полной вероятности): $\forall B$ верно $\sum_{i=1}^{k} P(A_i|B) = 1$, откуда

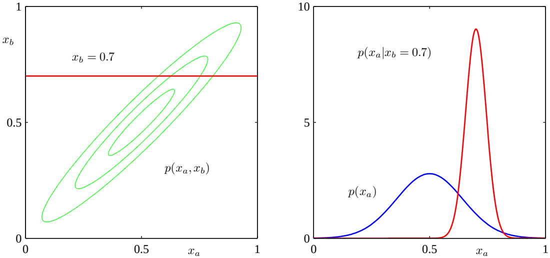
$$\sum_{i=1}^{k} \frac{P(B|A_i)P(A_i)}{P(B)} = 1 \qquad P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

- В интегральной форме

$$p(b) = \int p(b,a)da = \int p(b|a)p(a)da$$

Условное и маргинальное распределения

- Плотность распределения интересующей нас компоненты x_a многомерной случайной величины (x_a, x_b) можно получить двумя способами в зависимости от имеющейся информации
- Если нам неизвестны значения остальных компонент, мы маргинализуем плотность по ним: $p(x_a) = \int p(x_a, x_b) dx_b$
- Если значения других компонент нам известны, то мы обуславливаем плотность по ним: $p(x_a|x_b) = \frac{p(x_a, x_b)}{p(x_b)}$



Правило произведения вероятностей

- Правило произведения гласит, что любую совместную плотность всегда можно разбить на множители

$$p(a,b) = p(a|b)p(b) \qquad P(A,B) = P(A|B)P(B)$$

- Аналогично для многомерных совместных распределений

$$p(a_1,\dots,a_n) =$$
$$p(a_1|a_2,\dots,a_n)p(a_2|a_3,\dots,a_n)\dots p(a_{n-1}|a_n)p(a_n)$$

- Можно показать (Jaynes, 1995), что правила суммирования и произведения вероятностей являются единственными возможными операциями, позволяющими рассматривать вероятности как промежуточную ступень между истиной и ложью

Априорные и апостериорные суждения

- Предположим, мы пытаемся изучить некоторое явление
- У нас имеются некоторые знания, полученные до (лат. a priori) наблюдений/эксперимента. Это может быть опыт прошлых наблюдений, какие-то модельные гипотезы, ожидания
- В процессе наблюдений эти знания подвергаются постепенному уточнению. После (лат. a posteriori) наблюдений/эксперимента у нас формируются новые знания о явлении
- Будем считать, что мы пытаемся оценить неизвестное значение величины θ посредством наблюдений некоторых ее косвенных характеристик $x|\theta$

Формула Байеса

- Знаменитая формула Байеса (1763 г.) устанавливает правила, по которым происходит преобразование знаний в процессе наблюдений
- Обозначим априорные знания о величине θ за $p(\theta)$
- В процессе наблюдений мы получаем серию значений $\boldsymbol{x} = (x_1, \dots, x_n)$. При разных θ наблюдение выборки $\boldsymbol{x}$ более или менее вероятно и определяется значением правдоподобия $p(\boldsymbol{x}|\theta)$
- За счет наблюдений наши представления о значении θ меняются согласно формуле Байеса

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{\int p(\boldsymbol{x}|\theta)p(\theta)d\theta}$$

- Заметим, что знаменатель не зависит от θ и нужен исключительно для нормировки апостериорной плотности

Точечные оценки при использовании метода Байеса

- Хотя строгое применение байесовского оценивания подразумевает работу со всем апостериорным распределением, на практике его часто заменяют на точечную оценку, взятую в точке максимума апостериорной плотности

$$\hat{\theta}_{MP} = \arg\max P(\theta|\boldsymbol{x}) = \arg\max P(\boldsymbol{x}|\theta)P(\theta) =$$

$$\arg\max\left(\log P(\boldsymbol{x}|\theta) + \log P(\theta)\right)$$

- Это фактически регуляризация метода максимального правдоподобия, обспечивающая компромисс между априорными ограничениями и наблюдениями

План

Условная независимость случайных величин

- Случайные величины x и y называются условно независимыми от z, если

$$p(x,y|z) = p(x|z)p(y|z)$$

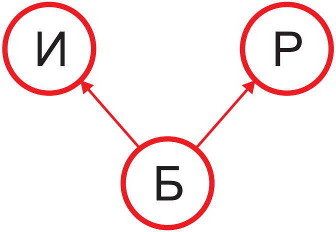
- Другими словами вся информация о взаимозависимостях между x и y содержится в z
- Заметим, что из безусловной независимости не следует условная и наоборот
- Основное свойство условно независимых случайных величин

$$p(z|x,y) = \frac{p(x,y|z)p(z)}{p(x,y)} = \frac{p(x|z)p(y|z)p(z)}{p(x,y)} =$$

$$\frac{p(x|z)p(z)p(y|z)p(z)}{p(x,y)p(z)} = \frac{p(z|x)p(z|y)}{p(z)p(x)p(y)p(x,y)} = \frac{1}{Z}\frac{p(z|x)p(z|y)}{p(z)}$$

Пример

- Рассмотрим следующую гипотетическую ситуацию: римские легионы во главе с императором атакуют вторгшихся варваров
- События «гибель императора» и «уничтожение Рима» не являются независимыми
- Однако, если нам дополнительно известен исход битвы с варварами, эти два события становятся независимыми
- В самом деле, если легионы битву проиграли, то судьба Рима мало зависит от того, был ли император убит в сражении



План

Классическая задача машинного обучения

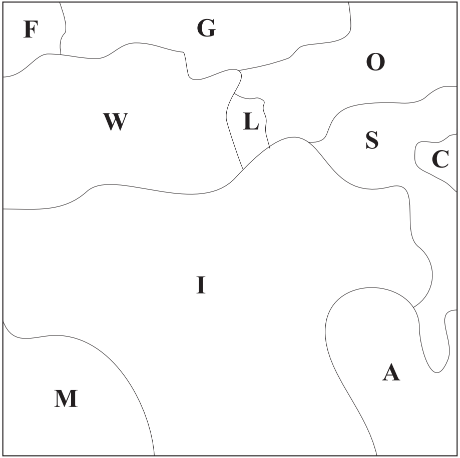
- Задачу машинного обучения можно трактовать как восстановление неизвестных зависимостей между наблюдаемымми переменными X и скрытыми (латентными) переменными T. В случае обучения с учителем такое восстановление производится по обучающей выборке Y
- В классических задачах машинного обучения предполагается, что обучающая выборка сформирована из однородных и независимых объектов $Y = \{(\boldsymbol{x}_i, \boldsymbol{t}_i)\}_{i=1}^n$
- До недавнего времени вероятностные методы обработки данных ограничивались только таким простейшим случаем, а изложение каждого метода начиналось со слов «Предположим, что нам дана выборка из независимых одинаково распределенных случайных величин...»

Задачи со структурными ограничениями

- Во многих задачах взаимосвязи между наблюдаемыми и скрытыми переменными носят сложный характер
- В частности, между отдельными переменными существуют вероятностные зависимости
- Факт зависимости переменных друг от друга удобно отображать с помощью неориентированного графа (марковской сети)
- Если связи между переменными причинно-следственные, то их удобно отображать в виде ориентированных графов (байесовских сетей)
- Основным средством работы с графическими моделями служит аппарат теории вероятностей, в ее байесовской интерпретации

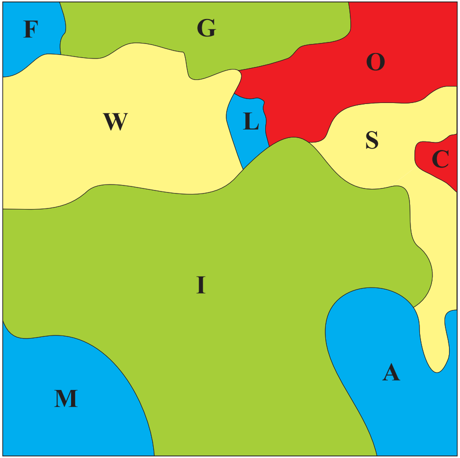
Простой пример

Задача о раскраске областей на плоскости так, чтобы никакие соседние не были окрашены в одинаковый цвет



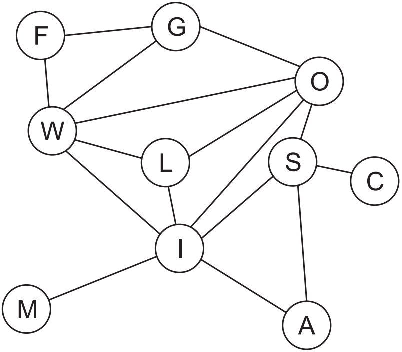
Простой пример

Задача о раскраске областей на плоскости так, чтобы никакие соседние не были окрашены в одинаковый цвет



Простой пример

Такая задача легко формулируется в терминах графической модели, в которой каждая вершина графа может находиться в одном из четырех состояний

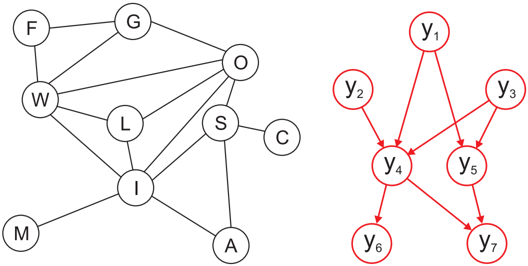


Вопрос залу: почему четырех?

План

Графические модели

- Графическая модель представляет собой ориентированный или неориентированный граф
- Вершины графа соответствуют переменным
- Ребра графа соответствуют вероятностным отношениям, определяющим непосредственные зависимости



Главные задачи в анализе графических моделей

Обозначим совокупность наблюдаемых переменных X, а ненаблюдаемых переменных T. Основными задачами в анализе графических моделей являются

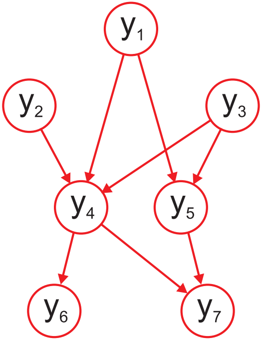
- Подсчет условного распределения на значения отдельной скрытой переменной $p(t_i|X)$–?
- Нахождение наиболее вероятной конфигурации скрытых переменных $p(T|X) \to \max_T$
- Оценка адекватности выбранной графической модели данным $p(X)$–?

Трудности, возникающие при использовании графических моделей

- Не во всех случаях существуют строгие алгоритмы вывода и обучения графических моделей
- Даже там, где они существуют, их применение может оказаться невозможно из-за высоких вычислительных требований и требований к памяти
- В настоящее время в мире активно разрабатываются приближенные эффективные методы обучения и принятия решения в графических моделях
  - Monte Carlo Markov chains (лекция 9)
  - Variational bounds (лекция 5)
  - Expectation propagation (лекция 5)
  - Loopy belief propagation (лекция 4)
  - Tree reweighted message passing (лекция 5))
  - и др.

План

Совместное распределение переменных



Совместное распределение системы переменных задается выражением

$$p(Y) = p(y_1, y_2, y_3, y_4, y_5, y_6, y_7) =$$
$$p(y_1)p(y_2)p(y_3)p(y_4|y_1, y_2, y_3)p(y_5|y_1, y_3)p(y_6|y_4)p(y_7|y_4, y_5).$$

Совместное и условные распределения

- В общем случае совместное распределение для ориентированного графа с n вершинами

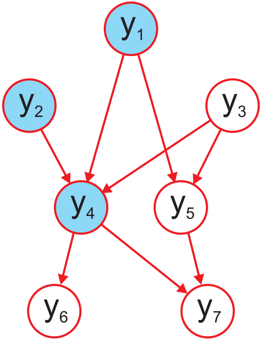
$$p(Y)=\prod_{i=1}^{n}p(y_i|\mathrm{pa}_i),$$

где pa_i — множество вершин-родителей y_i

- Обычно предполагается, что атомарные условные распределения $p(y_i|\mathrm{pa}_i)$ известны
- Зная атомарные распределения, мы можем рассчитать (хотя бы теоретически) любые условные вероятности одних подмножеств переменных по другим подмножествам переменных

Вычисление условных распределений I

- Вернемся к иллюстрации графической модели из семи переменных
- Пусть нам необходимо найти распределение (y_5, y_7) при заданных значениях y_1, y_2, y_4 и неизвестных y_3, y_6



Вычисление условных распределений II

- По определению условной вероятности

$$p(y_5, y_7 | y_1, y_2, y_4) = \frac{p(y_1, y_2, y_4, y_5, y_7)}{p(y_1, y_2, y_4)}$$

- Расписываем знаменатель

$$p(y_1, y_2, y_4) = p(y_1)p(y_2)p(y_4|y_1, y_2) = \{Sum\ rule\}$$
$$p(y_1)p(y_2) \int p(y_4|y_1, y_2, y_3)p(y_3)dy_3$$

- Аналогично числитель

$$p(y_1, y_2, y_4, y_5, y_7) = p(y_1)p(y_2)p(y_4|y_1, y_2)p(y_5|y_1)p(y_7|y_5, y_4) = p(y_1)\times$$
$$p(y_2)\left(\int p(y_4|y_1, y_2, y_3)p(y_3)dy_3\right)\left(\int p(y_5|y_1, y_3)p(y_3)dy_3\right)p(y_7|y_5, y_4)$$

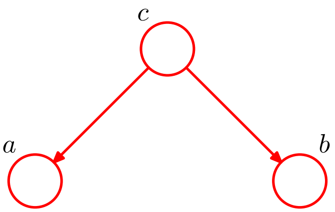
- Для взятия возникающих интегралов обычно пользуются различными аппроксимационными методами
- Таким образом, условное распределение выражено через известные атомарные распределения вида $p(y_i|\mathrm{pa}_i)$

План

Особенности использования байесовских сетей

- По смыслу построения байесовские сети не могут содержать ориентированные циклы, т.к. это будет нарушать правило умножения вероятностей
- Главным достоинством графических моделей является относительно простое выделение условно-независимых величин, которое облегчает дальнейший анализ, позволяя значительно уменьшить количество факторов, влияющих на данную переменную
- В байесовских сетях сделать это несколько сложнее, чем в марковских

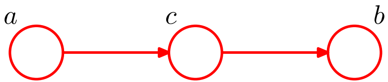
Граф 1



- Аналогия: Рим (a), император (b) и варвары (c)
- Переменные a и b независимы при заданном c
- Возможна маргинализация (исключение переменной)

$$p(a,b) = \int p(a|c)p(b|c)p(c)dc \neq p(a)p(b)$$

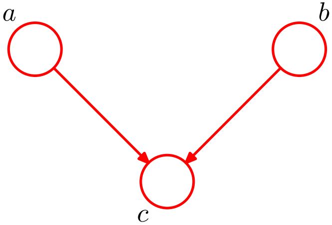
Граф 2



- Аналогия: хорошая работа (a), премия (c), яхта (b)
- Переменные a и b независимы при заданном c
- Возможна маргинализация (исключение переменной)

$$p(a,b) = p(a)\int p(b|c)p(c|a)dc \neq p(a)p(b)$$

Граф 3

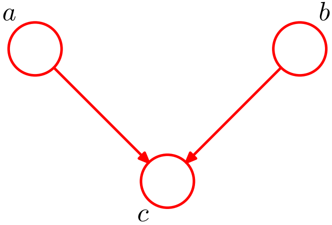


- Аналогия: вор (a), землетрясение (b) и сигнализация (c)
- Переменные a и b независимы, т.е. $p(a, b) = p(a)p(b)$, но не условно независимы!
- Зависимость $p(c|a, b)$ не может быть выражена через $p(c|a)$ и $p(c|b)$, хотя обратное верно

$$p(c|a) = \int p(c|a, b)p(b)db$$

План

Пример нестрогих вероятностных рассуждений I



- Рассмотрим последний граф подробнее. Введем обозначения событий: «сигнализация сработала/не сработала» ($s/\neg s$), «вор есть/вора нет» ($v/\neg v$) и «землетрясение произошло/не произошло» ($z/\neg z$)
- Пусть $p(s|v,\neg z) = p(s|v,z) = 1$, $p(s|\neg v,z) = 0.1$, $p(s|\neg v,\neg z) = 0$, $p(v) = 2\times 10^{-4}$, $p(z) = 10^{-2}$. Графическая модель полностью определена

Пример нестрогих вероятностных рассуждений II

Допустим, мы получили сигнал тревоги, т.е. имеет место событие s. Необходимо оценить вероятность того, что в квартире вор $p(v|s)$

$$p(v|s) = \frac{1}{Z}p(s|v)p(v) = \frac{p(s|v)p(v)}{p(s|v)p(v) + p(s|\neg v)p(\neg v)}$$

$$p(s|\neg v) = p(s|\neg v, z)p(z) + p(s|\neg v, \neg z)p(\neg z) = 10^{-3}$$

$$p(s|v) = 1$$

$$p(v|s) \approx \frac{1}{6}, \quad p(\neg v|s) \approx \frac{5}{6}, \quad Z \approx 1.2 \times 10^{-3}$$

Пример нестрогих вероятностных рассуждений III

- Пусть теперь дополнительно стало известно, что произошло землетрясение (событие z). Как изменится вероятность того, что в квартире вор $p(v|s,z)$?

$$p(v|s,z)=\frac{1}{Z}p(s|v,z)p(v|z),\quad p(v|z)=p(v)$$

$$Z=p(s|v,z)p(v|z)+p(s|\neg v,z)p(\neg v|z)=$$

$$1\times 2\times 10^{-4}+0.1\times(1-2\times 10^{-4})=0.1002$$

$$p(v|s,z)=0.002,\quad p(\neg v|s,z)=0.998$$

- Заметим, что события z и v перестали быть независимыми, и добавление сведений о значении z меняет знания о значении v. Это называется эффектом оправдания (explaining away)

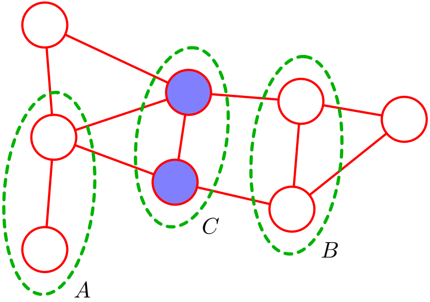
Примеры байесовских сетей

- Скрытые марковские модели (лекции 6, 7)
- Фильтр Калмана (лекция 8)
- Экспертные системы
- Вероятностный PCA (лекция 11)
- Смеси экспертов
- Факторный анализ
- и др.

План

Неориентированные графические модели

- При использовании ориентированных графов определение условной независимости не очень просто
- В марковских сетях это проще. На рисунке A и B независимы при условии C
- Ребра графа связывают переменные, между которыми существуют непосредственные (а не опосредованные) зависимости



Факторизация в марковских сетях

- Пусть y_i и y_j независимы при условии, что все остальные переменные нам известны, т.е. $p(y_i, y_j|Y_{\{i,j\}}) = p(y_i|Y_{\{i,j\}})p(y_j|Y_{\{i,j\}})$
- Это означает, что y_i и y_j не соединены ребром (иначе не было бы условной независимости)
- Запишем совместное распределение и применим правило умножения вероятностей

$$p(Y) = p(y_i, y_j|Y_{\{i,j\}})p(Y_{\{i,j\}}) = p(y_i|Y_{\{i,j\}})p(y_j|Y_{\{i,j\}})p(Y_{\{i,j\}})$$

- Таким образом, переменные, не соединенные ребрами, входят в разные множители совместного распределения

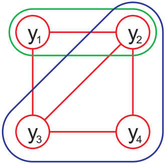
Потенциалы марковской сети

- В общем виде совместное распределение значений элементов сети записывается с помощью неотрицательных потенциальных функций, определенных на максимальных кликах

$$p(Y) = \frac{1}{Z}\prod_C \psi_C(Y_C), \quad Z = \sum_Y \prod_C \psi_C(Y_C), \quad \psi_C(Y_C) \geq 0$$

- На рисунке синяя клика является максимальной, а зеленая — нет. Совместное распределение имеет вид

$$p(Y) = \frac{1}{Z}\psi_1(y_1, y_2, y_3)\psi_2(y_2, y_3, y_4)$$



Потенциальная функция не обязана иметь вероятностную природу, но чем она больше, тем более вероятны соответствующие значения переменных. Обычно потенциальные функции задаются пользователем исходя из априорных предпочтений тех или иных конфигураций переменных. Реже — настраиваются по данным

Энергетическая нотация

- Иногда удобно ввести обозначение $\psi_C(Y_C) = \exp(-E_C(Y_C))$, где $E_C(Y_C)$ имеет смысл энергии
- Тогда задача нахождения наиболее вероятного состояния системы сводится к задаче минимизации полной энергии системы

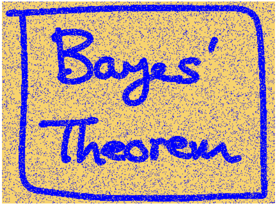
$$\arg\max p(Y) = \arg\max \frac{1}{Z} \prod_C \psi_C(Y_C) =$$

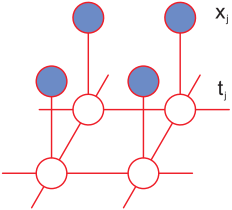
$$\arg\max \exp\left(-\sum_C E_C(Y_C)\right) = \arg\min \sum_C E_C(Y_C)$$

- Заметим, что в отличие от байесовских сетей для полного задания графической модели необходимо знать (или конструктивно уметь подсчитывать) нормировочную константу Z

План

Фильтрация изображений





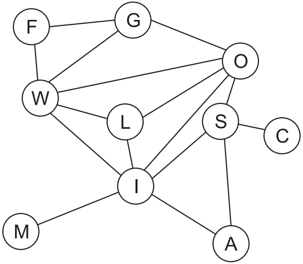
- Рассмотрим задачу фильтрации изображения. Пусть $x_i \in \{-1, 1\}$ — наблюдаемые пиксели бинарного изображения, а $t_i \in \{-1, 1\}$ — истинные значения пикселей
- Введем энергию системы

$$E(X, Y) = h\sum_i t_i - \beta \sum_{(i,j)\in E} t_i t_j - \eta \sum_i t_i x_i,$$

где $h \in \mathbb{R}$ позволяет отразить априорные предпочтения в пользу того или иного цвета (например, указать, что желтый цвет встречается чаще, чем синий), $\beta > 0$ выражает степень зависимости между соседними пикселями, а $\eta > 0$ показывает интенсивность шума

Разметка областей I

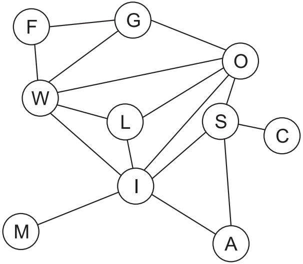
Вернемся к примеру со странами



Совместная плотность задается формулой

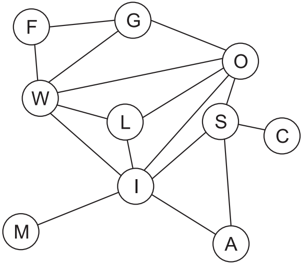
$$p(X) = \frac{1}{Z}\psi_1(F,G,W)\psi_2(G,O,W)\psi_3(W,O,L,I)\psi_4(I,S,O)\times$$
$$\times\ \psi_5(S,I,A)\psi_6(S,C)\psi_7(M,I)$$

Разметка областей II



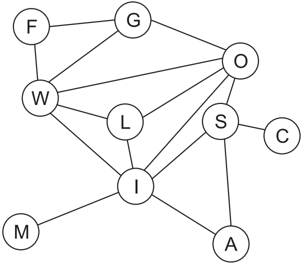
- Предположим, что переменные могут принимать одно из четырех значений $\{red, yellow, blue, green\}$
- Требование несовпадающих цветов областей эквивалентно условию равенства нулю потенциала, если хотя бы два его аргумента имеют одинаковое значение, например $\psi_5(red, blue, red) = 0$

Разметка областей III



Мы можем снизить число нежелаемых цветовых переходов (например, из желтого в красный) снизив соответствующие значения потенциалов $\psi_7(red, yellow)$, $\psi_7(yellow, red)$, $\psi_6(red, yellow)$ и $\psi_6(yellow, red)$

Разметка областей IV



Мы можем искусственно способствовать окраске отдельных регионов в выбранные цвета, вводя индивидуальные множители, например

$$\psi_1(F,G,W) = \phi_1(F,G,W)\phi_2(F)\phi_3(W).$$

Теперь можно увеличить значение $\phi_2(blue)$ и $\phi_3(green)$, чтобы получить политическую карту, привычную российскому глазу

Примеры марковских сетей

- Изображения (лекция 3)
- Социальные сети
- Случайные поля
- Карты сайтов
- и др.

План

Марковские vs. Байесовские сети

Сходства и различия двух типов графических моделей

Свойство	Марковские сети	Байесовские сети
Форма	Произв. потенциалов	Произв. потенциалов
Потенциалы	Произвольные	Усл. вероятности
Циклы	Разрешены	Запрещены
Нормировка	$Z=?$	$Z=1$
Условная нез-ть	Легко проверяема	Сложнее
Полнота	нет	нет
Анализ	MCMC, BP, и т.д..	Сводить к марковским

Существующее положение дел

- На сегодняшний день существуют эффективные алгоритмы (sum-product, max-product) анализа ациклических графов (деревьев), решающие все три основные задачи анализа графических моделей
- Частным случаем деревьев являются графы-цепочки, характеризующие, например, сигналы во времени
- В случае наличия циклов сложность точных алгоритмов резко возрастает
- Для анализа графов с циклами в основном используются приближенные методы (loopy BP, EP, MCMC)
- В некоторых частных случаях для анализа циклических сетей существуют эффективные точные алгоритмы, например, разрезы графов (лекция 3)

Сведение байесовских сетей к марковским

- Наиболее разработаны в настоящее время методы анализа марковских сетей
- Байесовскую сеть можно легко свести к марковской с потерей информации, переженив всех родителей (морализация)
- Заметим, что в приведенном примере все полезные свойства оказались потеряны, и мы получили клику, в которой все зависят от всех