ВАРИАЦИОННЫЕ ПРИБЛИЖЕНИЯ

Сергей Николенко

СПбГУ — Санкт-Петербург

8 апреля 2024 г.

Random facts:

- 8 апреля 1546 г. Тридентский собор признал Biblia Vulgata в качестве официальной версии Библии; работа над переводом на латынь шла ещё при папе Дамасии в IV-V веках, а в 1456 г. Иоганн Гутенберг воспроизвёл её текст в первой печатной книге
- 8 апреля 1513 г. Хуан Понсе де Леон объявил открытую им Флориду владением Испании (на самом деле он искал источник вечной молодости); в его честь назван Понсе, второй крупнейший город Пуэрто-Рико, и по некоторым подсчётам 30% современного населения Пуэрто-Рико являются дальними потомками Понсе де Леона
- 8 апреля 1820 г. Йоргос Кентротас, возделывая своё поле на Милосе, обнаружил разбитую на две части статую Венеры: торс без рук и задрапированную нижнюю часть
- 8 апреля 1912 г. прошёл премьерный показ кукольного мультфильма «Прекрасная Люканида, или Война усачей с рогачами»; Владислав Старевич создал любовную историю жука-рогача Люканиды и жука-усача Героса при помощи покадровой съёмки, но многие критики посчитали мультфильм результатом дрессировки жуков
- 8 апреля 1971 г. вблизи Лондона состоялся первый Всемирный конгресс цыган, который принял цыганские гимн и флаг; с тех пор 8 апреля — Международный день цыган

Вариационные приближения

Вариационный вывод

- Вариационный вывод: функционалы, производные по функциям... в общем, можно оптимизировать функционалы.
- Для нас это значит, что можно оптимизировать приближение q из какого-то класса к заданному p.
- Пусть есть $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ и $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$.
- Мы знаем $p(\mathbf{X}, \mathbf{Z})$ из модели, хотим найти приближение для $p(\mathbf{Z} \mid \mathbf{X})$ и $p(\mathbf{X})$.

Вариационный вывод

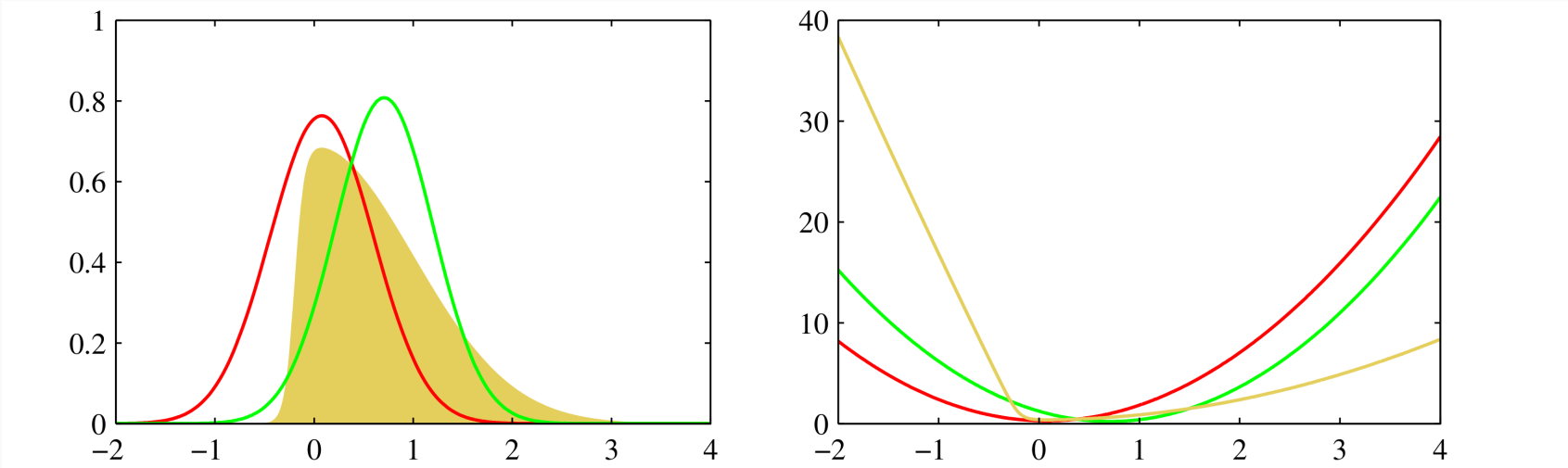
- Как и в EM:

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p), \text{ где}$$
$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z},$$
$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \frac{p(\mathbf{Z} \mid \mathbf{X})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$

- $\mathcal{L}(q)$ — это вариационная нижняя оценка, её можно теперь максимизировать, и KL будет автоматически минимизироваться.

Вариационный вывод

- Пример – сравним с лапласовским:



- Если $q(Z)$ произвольное, то мы просто получим $q(Z) = p(\mathbf{Z} \mid \mathbf{X})$; но это вряд ли получится.
- Будем ограничивать.

Факторизуемые распределения

- Главный частный случай — пусть $\mathbf{Z} = \mathbf{Z}_1 \sqcup \ldots \sqcup \mathbf{Z}_M$, и

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i).$$

- Но больше никаких предположений! В этом прелесть — оптимизируем сразу функции!
- Это соответствует теории среднего поля в физике (mean field theory).

Факторизуемые распределения

- Тогда:

$$\begin{aligned}\mathcal{L}(q) &= \int \prod_i q_i \left(\ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right) \mathrm{d}\mathbf{Z} \\ &= \int q_j \left[\int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \mathrm{d}\mathbf{Z}_i \right] \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \mathrm{d}\mathbf{Z}_j + \text{const} \\ &= \int q_j \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \mathrm{d}\mathbf{Z}_j + \text{const},\end{aligned}$$

$$\text{где } \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathrm{E}_{i \neq j} \left[\ln p(\mathbf{X}, \mathbf{Z}) \right] + \text{const}.$$

- Как максимизировать теперь $\mathcal{L}(q)$ по q_j?

Факторизуемые распределения

- Надо заметить, что мы получили там KL-дивергенцию между $q_j(\mathbf{Z}_j)$ и $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$.
- Т.е. оптимальное решение получится при

$$\ln q_j^*(\mathbf{Z}_j) = \mathrm{E}\left[\ln p(\mathbf{X}, \mathbf{Z})\right] + \mathrm{const}.$$

- Константа здесь просто для нормализации.
- Оказывается, достаточно взять ожидание от логарифма совместного распределения.
- Но явных формул не получается, потому что ожидание считается по остальным q_i^*, $i \neq j$.
- И всё-таки обычно что-то можно сделать; давайте рассмотрим примеры.

Двумерный гауссиан

- Первый пример — приблизим двумерный гауссиан одномерными:

$$p(\mathbf{z}) = N(\mathbf{z} \mid \mu, \Lambda^{-1}),$$

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}.$$

- И мы хотим приблизить $q(\mathbf{z}) = q_1(z_1)q_2(z_2)$.
- Повычисляем...

Двумерный гауссиан

- ...получится, что

$$\ln q_1^*(z_1) = -\frac{1}{2} z_1^2 \Lambda_{11} + z_1 \mu_{11} \Lambda_{11} - z_1 \Lambda_{12} (\mathrm{E}[z_2] - \mu_2) + \mathrm{const}.$$

- Чудесным образом получился гауссиан! Сам собой, без предположений.
- Найдём его среднее и дисперсию...

Двумерный гауссиан

- ...получится

$$q_1^*(z_1) = N(z_1 \mid m_1, \Lambda_{11}^{-1}), \text{ где}$$

$$m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathrm{E}[z_2] - \mu_2).$$

- Аналогично,

$$q_2^*(z_2) = N(z_2 \mid m_2, \Lambda_{11}^{-1}), \text{ где}$$

$$m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathrm{E}[z_1] - \mu_1).$$

- Какое решение у этой системы?

Двумерный гауссиан

- Да просто $\mathrm{E}[z_1] = m_1 = \mu_1$, $\mathrm{E}[z_2] = m_2 = \mu_2$.
- А если бы мы минимизировали $\mathrm{KL}(p\|q)$, получилось бы

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{Z}) \left[\sum_i \ln q_i(\mathbf{Z}_i)\right] \mathrm{d}\mathbf{Z} + \mathrm{const},$$

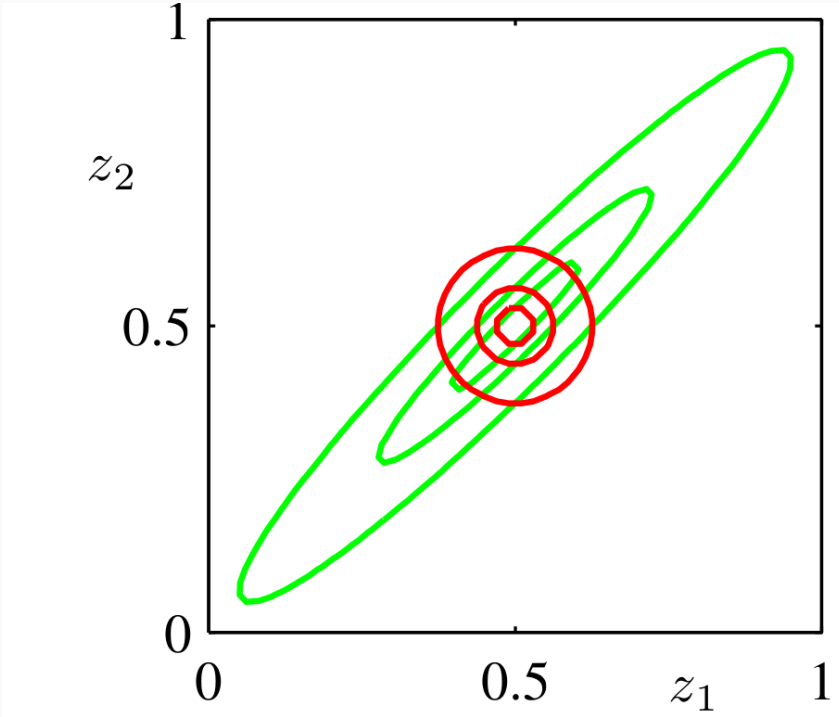
  и можно оптимизировать по отдельности:

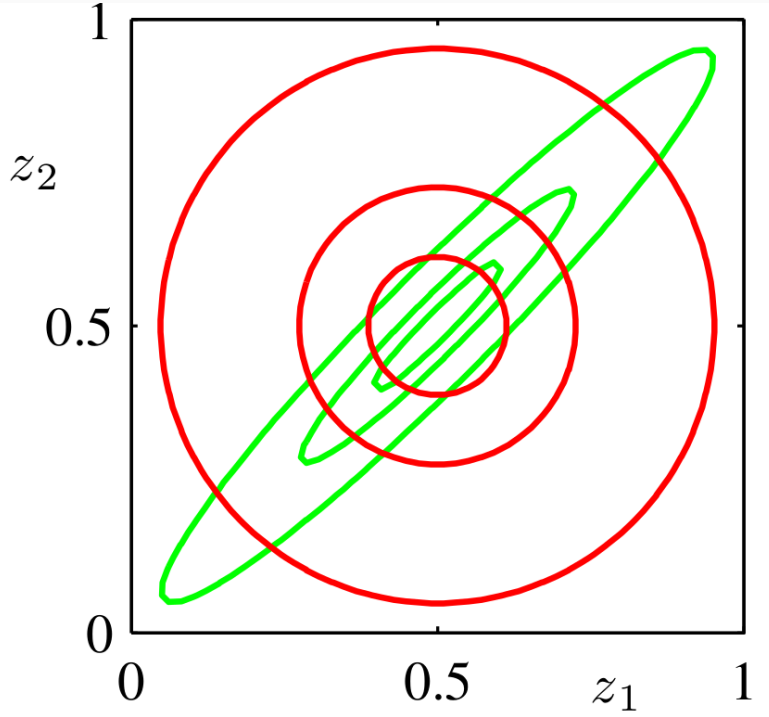
$$q_j^*(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} \mathrm{d}\mathbf{Z}_i = p(\mathbf{Z}_j).$$

- Т.е. просто маргинализация.
- Почему бы так и не сделать? В чём разница?

Разные KL-дивергенции

- Разные дисперсии ответа:





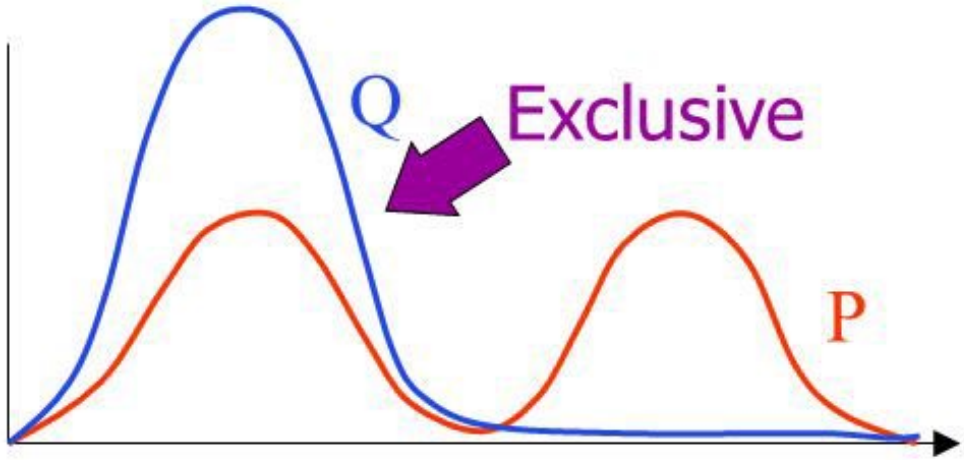
Разные KL-дивергенции

- Минимизация $\mathrm{KL}(p\|q)$ накрывает всю p, а $\mathrm{KL}(q\|p)$ строит всю q в регионе больших p:

Minimising

$\mathrm{KL}(Q||P)$

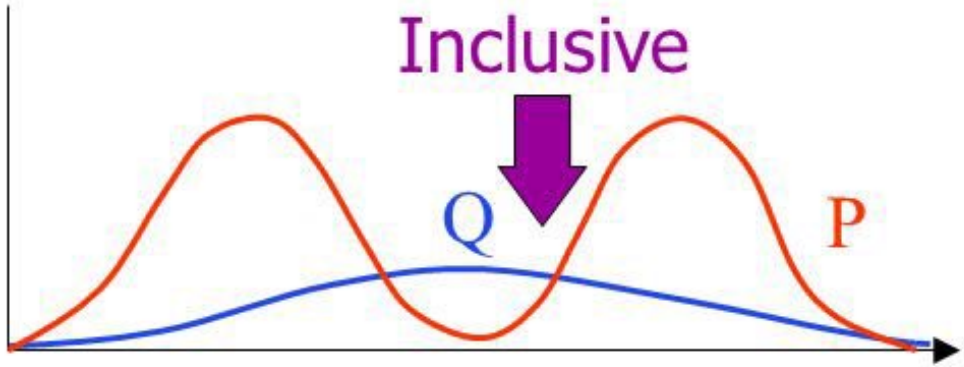
$$= \sum_H Q(H) \ln \frac{Q(H)}{P(H|V)}$$



Minimising

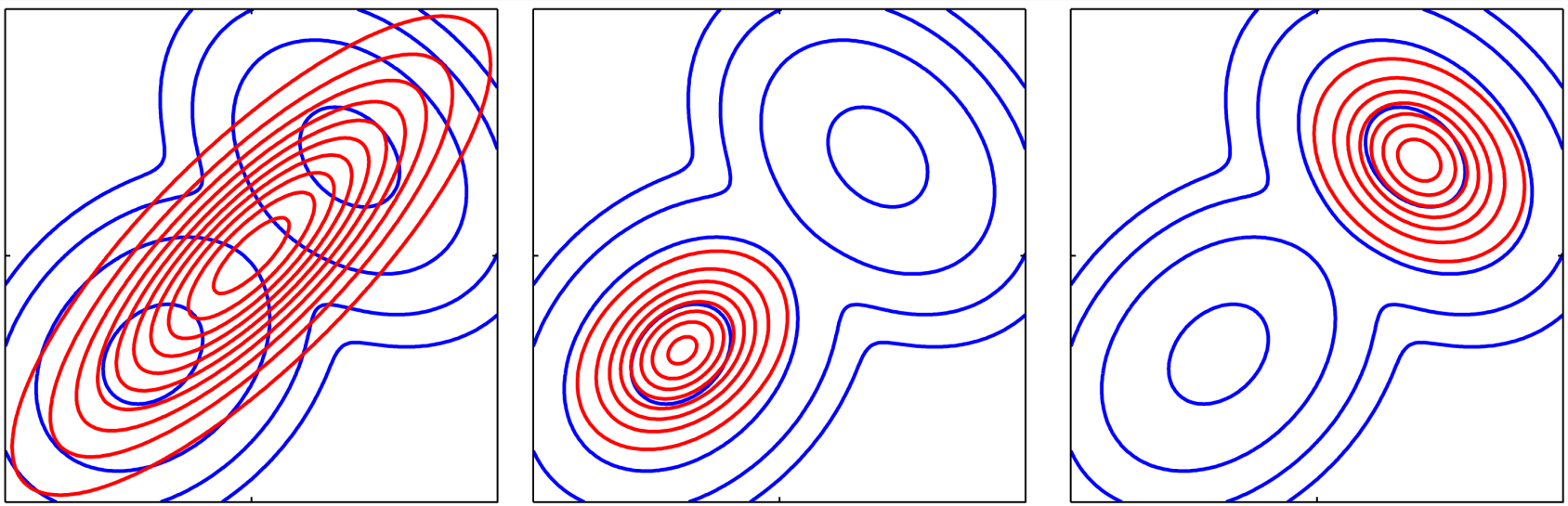
$\mathrm{KL}(P||Q)$

$$= \sum_H P(H|V) \ln \frac{P(H|V)}{Q(H)}$$



Разные KL-дивергенции

- Например, для двумерного гауссиана:



- В машинном обучении гораздо интереснее, конечно, пик найти.

Вариационное приближение для гауссиана

Одномерный гауссиан

- И ещё пример: давайте найдём параметры одномерного гауссиана по точкам $\mathbf{X} = \{x_1, \dots, x_N\}$. Правдоподобие:

$$p(\mathbf{X} \mid \mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} e^{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2}.$$

- Вводим сопряжённые априорные распределения:

$$p(\mu \mid \tau) = N(\mu \mid \mu_0, (\lambda_0\tau)^{-1}),$$
$$p(\tau) = \mathrm{Gamma}(\tau \mid a_0, b_0).$$

- Мы это только что подсчитали точно, но давайте приблизим теперь апостериорное распределение как

$$q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau).$$

- На самом деле так не раскладывается!
- Это то, что мы делали для $q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$. Посчитаем...

Одномерный гауссиан

- ...$q_\mu(\mu)$ – гауссиан с параметрами

$$\mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N}, \quad \lambda_N = (\lambda_0 + N)\mathrm{E}[\tau].$$

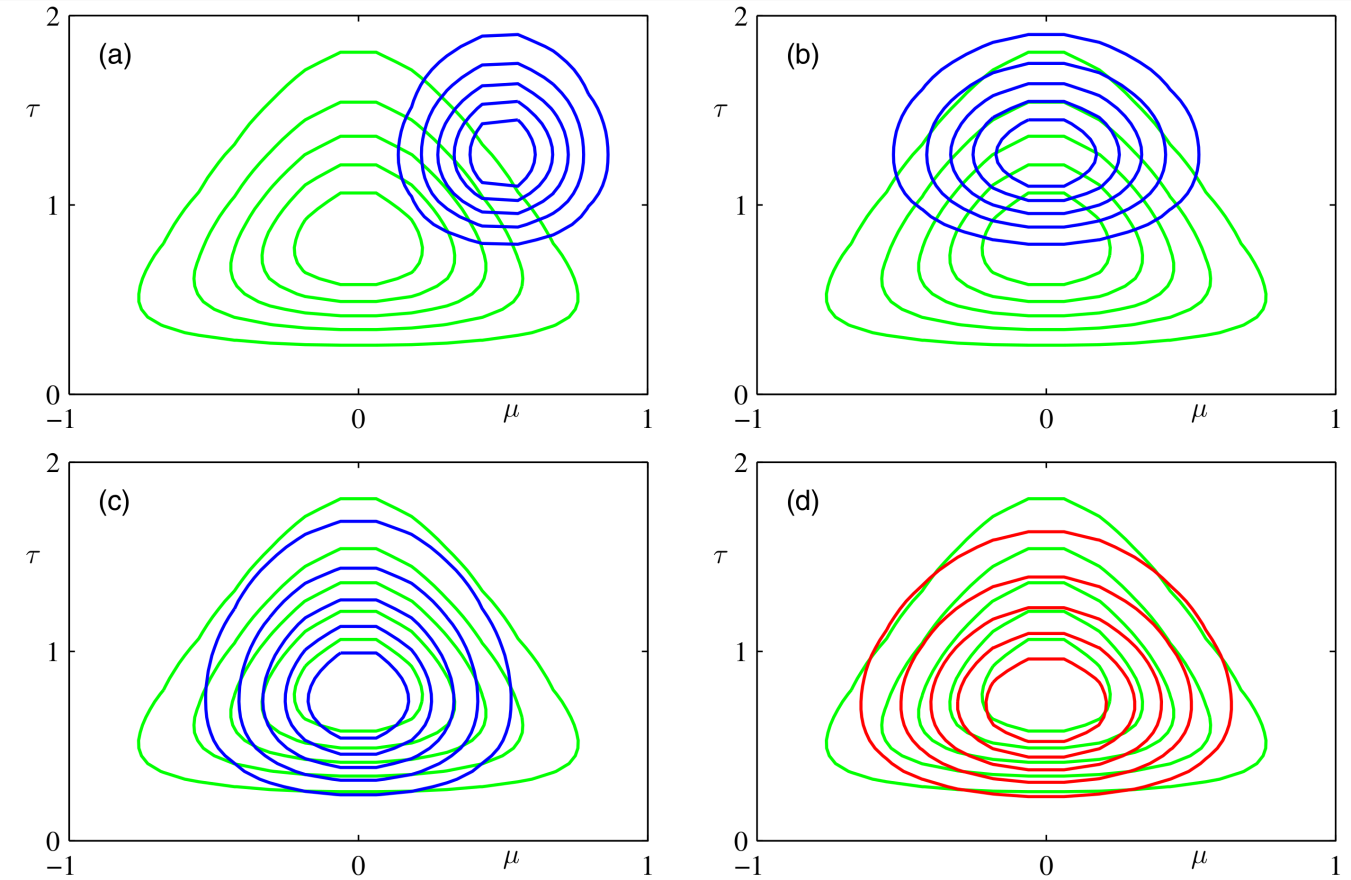
- А $q_\tau(\tau)$ – гамма-распределение с параметрами

$$a_N = a_0 + \frac{N}{2}, \quad b_N = b_0 + \frac{1}{2}\mathrm{E}_\mu\left[\sum_n (x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right].$$

- Всё получилось как надо, но без предположений о форме q_τ и q_μ.

Одномерный гауссиан

- Вот такой вывод в пространстве (μ, τ):



- А для $\mu_0 = a_0 = b_0 = \lambda_0 = 0$ (non-informative priors) можно и точно посчитать...

Одномерный гауссиан

- Получатся моменты для μ

$$\mathrm{E}[\mu] = \bar{x}, \quad \mathrm{E}[\mu^2] = \bar{x}^2 + \frac{1}{N\mathrm{E}[\tau]}.$$

- Это можно подставить и найти $\mathrm{E}[\tau]$:

$$\frac{1}{\mathrm{E}[\tau]} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2.$$

- Автоматически получили несмещённую оценку дисперсии!

Вариационное приближение для смеси гауссианов

Смесь гауссианов

- Смесь гауссианов: $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$,

$$p(\mathbf{Z} \mid \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}},$$

$$p(\mathbf{X} \mid \mathbf{Z}, \mu, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} N\left(\mathbf{x}_n \mid \mu_k, \Lambda_k^{-1}\right).$$

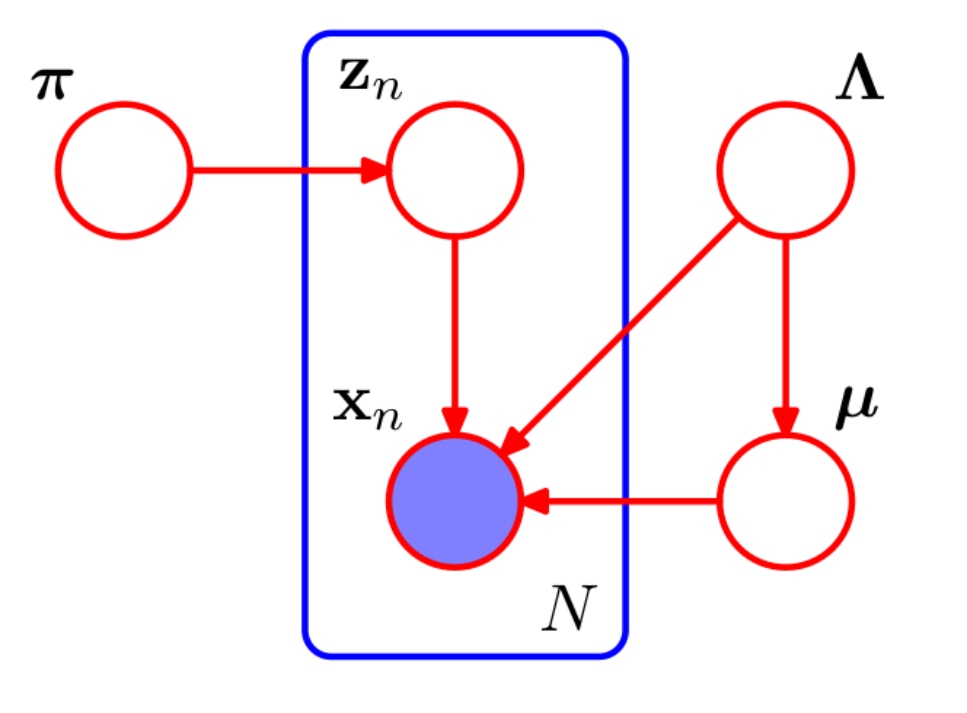
- Выберем сопряжённые априорные распределения:

$$p(\pi) = \mathrm{Dir}(\pi \mid \alpha_0) = C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1},$$

$$\begin{aligned} p(\mu, \Lambda) &= p(\mu \mid \Lambda) p(\Lambda) \\ &= \prod_{k=1}^{K} N\left(\mu_k \mid \mathbf{m}_0, (\beta_0 \Lambda_k)^{-1}\right) W(\Lambda_k \mid \mathbf{W}_0, \nu_0). \end{aligned}$$

Смесь гауссианов

- Вот такая графическая модель:



- Распределение Дирихле пусть будет симметричное для простоты; часто ещё $\mathbf{m}_0 = 0$.
- Заметьте разницу между латентными переменными и параметрами модели.

Вариационное приближение

- Теперь вариационное приближение. Сначала сама факторизация:

$$p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Lambda) = p(\mathbf{X} \mid \mathbf{Z}, \mu, \Lambda)p(\mathbf{Z} \mid \pi)p(\pi)p(\mu \mid \Lambda)p(\Lambda).$$

- Мы наблюдаем только $\mathbf{X}$, остальное всё надо как-то оценить.
- Интересно, что единственное предположение про наше вариационное приближение выглядит так:

$$q(\mathbf{Z}, \pi, \mu, \Lambda) = q(\mathbf{Z})q(\pi, \mu, \Lambda).$$

- И всё! Дальше всё само собой получится. Но не сразу...

Вариационное приближение

- Сначала $q^*(\mathbf{Z})$:

$$\begin{aligned}\ln q^*(\mathbf{Z}) &= \mathrm{E}_{\pi,\mu,\Lambda}\left[\ln p(\mathbf{X},\mathbf{Z},\pi,\mu,\Lambda)\right] + \text{const} \\ &= \mathrm{E}_{\pi,\mu,\Lambda}\left[\ln p(\mathbf{Z}\mid\pi)\right] + \mathrm{E}_{\mu,\Lambda}\left[\ln p(\mathbf{X}\mid\mathbf{Z},\mu,\Lambda)\right] + \text{const} \\ &= \ldots = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ln\rho_{nk} + \text{const},\end{aligned}$$

$$\text{где } \ln\rho_{nk} = \mathrm{E}[\ln\pi_k] + \frac{1}{2}\mathrm{E}[\ln|\Lambda_k|] - \frac{D}{2}\ln(2\pi) - \\ - \frac{1}{2}\mathrm{E}_{\mu_k,\Lambda_k}\left[(\mathbf{x}_n-\mu_k)^\top\Lambda_k(\mathbf{x}_n-\mu_k)\right].$$

- Нормируем:

$$q^*(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}}, \text{ где } r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K}\rho_{nj}}.$$

Вариационное приближение

- Теперь $\mathrm{E}[z_{nk}] = r_{nk}$, т.е. r_{nk} – то, насколько точка $\mathbf{x}_n$ принадлежит кластеру k.
- Можно определить статистики с их учётом, как обычно:

$$N_k = \sum_{n=1}^{N} r_{nk},$$

$$\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n,$$

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^\top.$$

- То же самое происходило и в EM-алгоритме.

Вариационное приближение

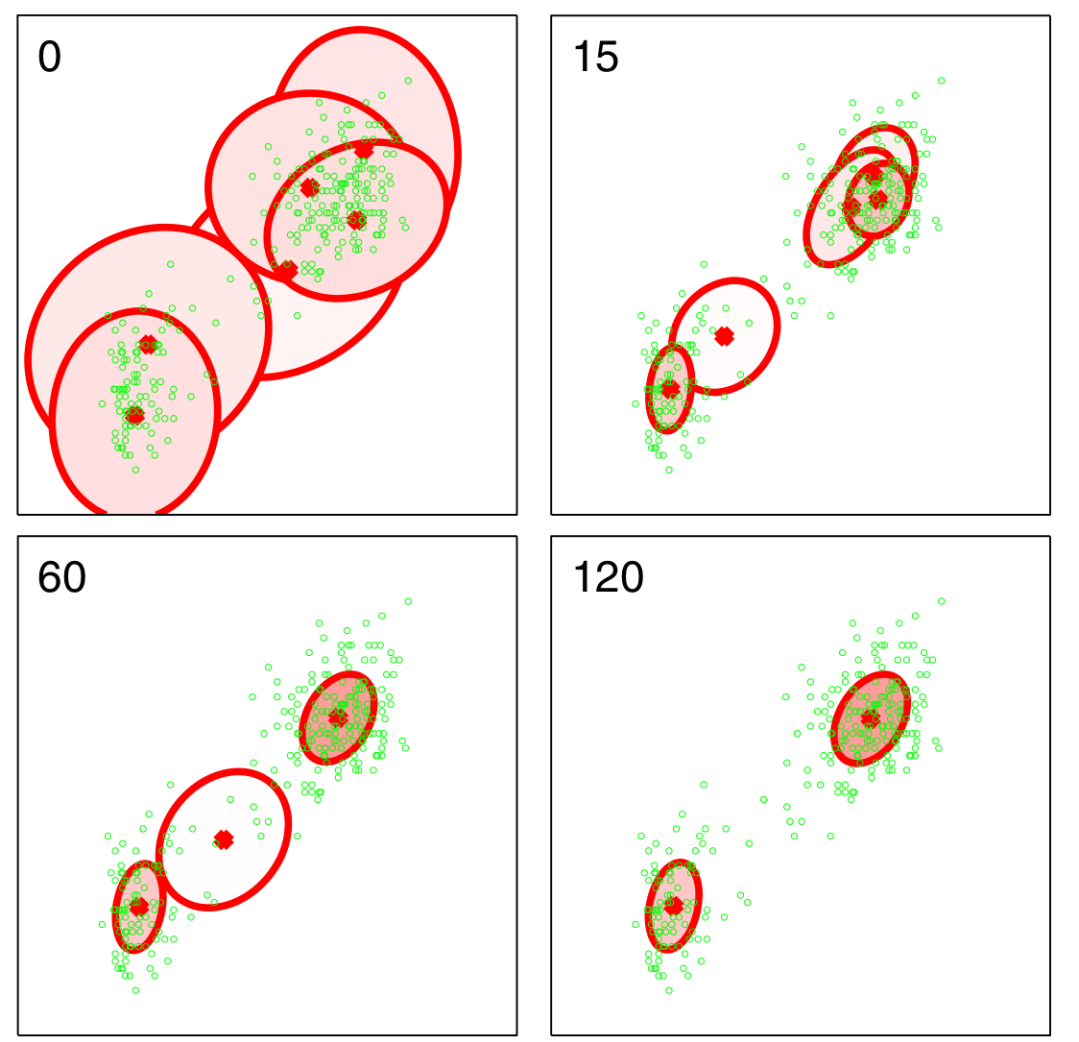
- Теперь $q^*(\pi, \mu, \Lambda)$:

$$\ln q^*(\pi, \mu, \Lambda) = \ln p(\pi) + \sum_{k=1}^{K} \ln p(\mu_k, \Lambda_k) + \mathrm{E}_{\mathbf{Z}}[\ln p(\mathbf{Z} \mid \pi)]$$
$$+ \sum_{k=1}^{K} \sum_{n=1}^{N} \mathrm{E}[z_{nk}] \ln N(\mathbf{x}_n \mid \mu_k \Lambda_k^{-1}) + \mathrm{const}.$$

- Вот уже получилось, что $q^*(\pi, \mu, \Lambda)$ раскладывается в $q^*(\pi)q^*(\mu, \Lambda)$, опять же без предположений.
- Более того, $q^*(\mu, \Lambda) = \prod_{k=1}^{K} q(\mu_k, \Lambda_k)$.
- И теперь можно по отдельности посчитать (упражнение), получится типичный M-шаг.
- Причём распределения останутся той же формы (т.к. были сопряжённые).

Вариационное приближение

- Теперь даже model selection автоматически получается, просто у некоторых компонент $N_k \approx 0$:



- Никакого оверфиттинга или коллапса компонент.

Другие примеры

- Есть другие примеры вариационных приближений.
- Обращение матриц; например, для линейной регрессии надо посчитать $\beta^* = C^{-1}\mathbf{b}$:

$$J(\beta) = \frac{1}{2}(\beta^* - \beta)^\top C(\beta^* - \beta) = \ldots = \text{Const} - \beta^\top \mathbf{b} + \frac{1}{2}\beta^\top C\beta,$$

  и теперь можно решать такую задачу выпуклой оптимизации.
- Метод конечных элементов – для уравнения Пуассона $-u''(x) = f(x)$, $x \in (a, b)$:

$$J(u) = \frac{1}{2}\int_a^b \left(u'(x) - u^{*\prime}(x)\right)^2 \mathrm{d}x = \ldots = \text{Const} - \int_a^b u(x)f(x)\mathrm{d}x + \frac{1}{2}\int_a^b u'(x)^2\mathrm{d}x,$$

  и если ищем в подпространстве $\tilde{u}(x) = \sum_{i=1}^k \alpha_i\phi_i(x)$, то опять

$$\tilde{J}(\alpha) = \alpha^\top \mathbf{b} + \frac{1}{2}\alpha^\top C\alpha.$$

В графических моделях

- В графических моделях – теория среднего поля (mean field theory). Пусть дано $p(\mathbf{x})$, $\mathbf{x} = (\mathbf{x}_v, \mathbf{x}_h)$, и надо найти

$$\log p(\mathbf{x}_v) = \log \sum_{\mathbf{x}_h} p(\mathbf{x}_v, \mathbf{x}_h), \quad p(\mathbf{x}_h \mid \mathbf{x}_v) = p(\mathbf{x}_h, \mathbf{x}_v)/p(\mathbf{x}_v).$$

- Опять делаем тот же трюк:

$$J(q) = \log p(\mathbf{x}_v) - \mathrm{KL}(q_{\mathbf{x}_h} \| p_{\mathbf{x}_h|\mathbf{x}_v}) = \log p(\mathbf{x}_v) - \sum_{\mathbf{x}_h} q(\mathbf{x}_h) \log \frac{q(\mathbf{x}_h)}{p(\mathbf{x}_h \mid \mathbf{x}_v)}$$

$$\ldots = H(q) + \mathbb{E}_q \left[\log p(\mathbf{x}_h, \mathbf{x}_v)\right] = H(q) + \sum_{C \in C} \sum_{\mathbf{x}_{C \cap h}} q(\mathbf{x}_{C \cap h}) \log \Psi_C(\mathbf{x}_C),$$

где $q(\mathbf{x}_{C \cap h})$ – маргинальная вероятность по скрытым переменным из клики C.

- Теория среднего поля – это когда $q(\mathbf{x}_h) = \prod_{i \in h} q_i(x_i)$.

Наивный байесовский классификатор

Категоризация текстов

- Классическая задача машинного обучения и information retrieval – категоризация текстов.
- Дан набор текстов, разделённый на категории. Нужно обучить модель и потом уметь категоризовать новые тексты.
- Атрибуты $a_1, a_2, \ldots, a_n$ – это слова, v – тема текста (или атрибут вроде «спам / не спам»).
- Bag-of-words model: забываем про порядок слов, составляем словарь. Теперь документ – это вектор, показывающий, сколько раз каждое слово из словаря в нём встречается.

Naive Bayes

- Заметим, что даже это – сильно упрощённый взгляд: для слов ещё довольно-таки важен порядок, в котором они идут...
- Но и это ещё не всё: получается, что $p(a_1, a_2, \dots, a_n | x = v)$ – это вероятность *в точности такого набора слов* в сообщениях на разные темы. Очевидно, такой статистики взять неоткуда.
- Значит, надо дальше делать упрощающие предположения.
- Наивный байесовский классификатор – самая простая такая модель: давайте предположим, что все слова в словаре условно независимы при условии данной категории.

Naive Bayes

- Иначе говоря:

$$p(a_1, a_2, \dots, a_n | x = v) = p(a_1 | x = v) p(a_2 | x = v) \dots p(a_n | x = v).$$

- Итак, наивный байесовский классификатор выбирает v как

$$v_{NB}(a_1, a_2, \dots, a_n) = \arg\max_{v \in V} p(x = v) \prod_{i=1}^{n} p(a_i | x = v).$$

- В парадигме классификации текстов мы предполагаем, что разные слова в тексте на одну и ту же тему появляются независимо друг от друга. Однако, несмотря на такие бредовые предположения, naive Bayes на практике работает очень даже неплохо (и этому есть разумные объяснения).

Многомерная модель

- Но в деталях реализации наивного байесовского классификатора есть тонкости.
- Сейчас мы рассмотрим два разных подхода к naive Bayes, которые дают разные результаты: мультиномиальный (multinomial) и многомерный (multivariate).

Многомерная модель

- В многомерной модели документ – это вектор бинарных атрибутов, показывающих, встретилось ли в документе то или иное слово.
- Когда мы подсчитываем правдоподобие документа, мы перемножаем вероятности того, что встретилось каждое слово из документа и вероятности того, что не встретилось каждое (словарное) слово, которое не встретилось.
- Получается модель многомерных испытаний Бернулли. Наивное предположение в том, что события «встретилось ли слово» предполагаются независимыми.

Многомерная модель

- Математически: пусть $V = \{w_t\}_{t=1}^{|V|}$ – словарь. Тогда документ d_i – это вектор длины $|V|$, состоящий из битов B_{it}; $B_{it} = 1$ iff слово w_t встречается в документе d_i.
- Правдоподобие принадлежности d_i классу c_j:

$$p(d_i \mid c_j) = \prod_{t=1}^{|V|} \left(B_{it} p(w_t \mid c_j) + (1 - B_{it})(1 - p(w_t \mid c_j))\right).$$

- Для обучения такого классификатора нужно обучить вероятности $p(w_t \mid c_j)$.

Многомерная модель

- Обучение – дело нехитрое: пусть дан набор документов $D = \{d_i\}_{i=1}^{|D|}$, которые уже распределены по классам c_j (возможно, даже вероятностно распределены), дан словарь $V = \{w_t\}_{t=1}^{|V|}$, и мы знаем биты B_{it} (знаем документы).
- Тогда можно подсчитать оптимальные оценки вероятностей того, что то или иное слово встречается в том или ином классе (при помощи лапласовой оценки):

$$p(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{|D|} B_{it} p(c_j \mid d_i)}{2 + \sum_{i=1}^{|D|} p(c_j \mid d_i)}.$$

Многомерная модель

- Априорные вероятности классов можно подсчитать как $p(c_j) = \frac{1}{|D|}\sum_{i=1}^{|D|} p(c_j \mid d_i)$.
- Тогда классификация будет происходить как

$$c = \arg\max_j p(c_j)p(d_i \mid c_j) =$$

$$= \arg\max_j \left(\frac{1}{|D|}\sum_{i=1}^{|D|} p(c_j \mid d_i)\right) \prod_{t=1}^{|V|} \left(B_{it}p(w_t \mid c_j) + (1 - B_{it})(1 - p(w_t \mid c_j))\right) =$$

$$= \arg\max_j \left(\log(\sum_{i=1}^{|D|} p(c_j \mid d_i)) + \sum_{t=1}^{|V|} \log\left(B_{it}p(w_t \mid c_j) + (1 - B_{it})(1 - p(w_t \mid c_j))\right)\right).$$

Мультиномиальная модель

- В мультиномиальной модели документ – это последовательность событий. Каждое событие – это случайный выбор одного слова из того самого «bag of words».
- Когда мы подсчитываем правдоподобие документа, мы перемножаем вероятности того, что мы достали из мешка те самые слова, которые встретились в документе. Наивное предположение в том, что мы достаём из мешка разные слова независимо друг от друга.
- Получается мультиномиальная генеративная модель, которая учитывает количество повторений каждого слова, но не учитывает, каких слов *нет* в документе.

Мультиномиальная модель

- Математически: пусть $V = \{w_t\}_{t=1}^{|V|}$ – словарь. Тогда документ d_i – это вектор длины $|d_i|$, состоящий из слов, каждое из которых «вынуто» из словаря с вероятностью $p(w_t \mid c_j)$.
- Правдоподобие принадлежности d_i классу c_j:

$$p(d_i \mid c_j) = p(|d_i|)|d_i|! \prod_{t=1}^{|V|} \frac{1}{N_{it}!} p(w_t \mid c_j)^{N_{it}},$$

  где N_{it} – количество вхождений w_t в d_i.
- Для обучения такого классификатора тоже нужно обучить вероятности $p(w_t \mid c_j)$.

Мультиномиальная модель

- Обучение: пусть дан набор документов $D = \{d_i\}_{i=1}^{|D|}$, которые уже распределены по классам c_j (возможно, даже вероятностно распределены), дан словарь $V = \{w_t\}_{t=1}^{|V|}$, и мы знаем вхождения N_{it}.
- Тогда можно подсчитать апостериорные оценки вероятностей того, что то или иное слово встречается в том или ином классе (не забываем сглаживание – правило Лапласа):

$$p(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{|D|} N_{it} p(c_j \mid d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{is} p(c_j \mid d_i)}.$$

Мультиномиальная модель

- Априорные вероятности классов можно подсчитать как $p(c_j) = \frac{1}{|D|}\sum_{i=1}^{|D|} p(c_j \mid d_i)$.
- Тогда классификация будет происходить как

$$c = \arg\max_j p(c_j)p(d_i \mid c_j) =$$

$$= \arg\max_j \left(\frac{1}{|D|}\sum_{i=1}^{|D|} p(c_j \mid d_i) \right) p(|d_i|)|d_i|! \prod_{t=1}^{|V|} \frac{1}{N_{it}!} p(w_t \mid c_j)^{N_{it}} =$$

$$= \arg\max_j \left(\log \left(\sum_{i=1}^{|D|} p(c_j \mid d_i) \right) + \sum_{t=1}^{|V|} N_{it} \log p(w_t \mid c_j) \right).$$

Как можно обобщить наивный байес

- В наивном байесе есть два сильно упрощающих дело предположения:
  - мы знаем метки тем всех документов;
  - у каждого документа только одна тема.
- Мы сейчас уберём оба эти ограничения.
- Во-первых, что можно сделать, если мы не знаем метки тем, т.е. если датасет неразмеченный?

Кластеризация

- Тогда это превращается в задачу кластеризации.
- Её можно решать EM-алгоритмом (Expectation-Maximization, используется в ситуациях, когда есть много скрытых переменных, причём если бы мы их знали, модель стала бы простой):
  - на E-шаге считаем ожидания того, какой документ какой теме принадлежит;
  - на M-шаге пересчитываем наивным байесом вероятности $p(w \mid t)$ при фиксированных метках.
- Это простое обобщение.

Спасибо за внимание!