

СКРЫТЫЕ МАРКОВСКИЕ МОДЕЛИ II

Сергей Николенко

СПбГУ — Санкт-Петербург

27 октября 2017 г.

Random facts:

- 27 октября 1962 г. --- «Чёрная суббота», день, когда мир был ближе всего к глобальной ядерной войне; в этот день американский разведывательный самолёт был сбит над Кубой советской зенитной ракетой
- 27 октября 1982 г. КНР объявила, что население Китая превысило миллиард
- 27 октября --- Всемирный день аудиовизуального наследия

ОБУЧАЕМ СКРЫТУЮ МАРКОВСКУЮ МОДЕЛЬ

ОБЩАЯ СУТЬ ТРЕТЬЕЙ ЗАДАЧИ

- Аналитически найти глобальный максимум $p(D|\lambda)$ у нас никак не получится.
- Зато мы рассмотрим итеративную процедуру (по сути — градиентный подъём), которая приведёт к локальному максимуму.
- Это называется алгоритм Баума–Велха (Baum–Welch algorithm). Он является на самом деле частным случаем алгоритма EM.

- Теперь нашими вспомогательными переменными будут вероятности того, что мы во время t в состоянии x_i , а во время $t + 1$ — в состоянии x_j :

$$\xi_t(i, j) = p(q_t = x_i, q_{t+1} = x_j | D, \lambda).$$

- Если переписать через уже знакомые переменные:

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(d_{t+1}) \beta_{t+1}(j)}{p(D | \lambda)} = \frac{\alpha_t(i) a_{ij} b_j(d_{t+1}) \beta_{t+1}(j)}{\sum_i \sum_j \alpha_t(i) a_{ij} b_j(d_{t+1}) \beta_{t+1}(j)}.$$

- Отметим также, что $\gamma_t(i) = \sum_j \xi_t(i, j)$.

- $\sum_t \gamma_t(i)$ — это ожидаемое количество переходов из состояния x_i , а $\sum_t \xi_t(i, j)$ — из x_i в x_j .
- Теперь на шаге M мы будем переоценивать вероятности:

$$\bar{\pi}_i = \text{ожидаемая частота в } x_i \text{ на шаге } 1 = \gamma_1(i),$$

$$\bar{a}_{ij} = \frac{\text{к-во переходов из } x_i \text{ в } x_j}{\text{к-во переходов из } x_i} = \frac{\sum_t \xi_t(i, j)}{\sum_t \gamma_t(i)}.$$

$$\bar{b}_j(k) = \frac{\text{к-во появлений в } x_i \text{ и наблюдений } v_k}{\text{к-во появлений в } x_i} = \frac{\sum_{t:d_t=v_k} \gamma_t(i)}{\sum_t \gamma_t(i)}.$$

- EM-алгоритм приведёт к цели: начать с $\lambda = (A, B, \pi)$, подсчитать $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$, снова пересчитать параметры и т.д.

- Kullback–Leibler distance (divergence) — это информационно-теоретическая мера того, насколько далеки распределения друг от друга.

$$D_{KL}(p_1, p_2) = \sum_x p_1(x) \log \frac{p_1(x)}{p_2(x)}.$$

- Известно, что это расстояние всегда неотрицательно, равно нулю iff $p_1 \equiv p_2$.

- Мы определим

$$p_1(Q) = \frac{p(Q, D|\lambda)}{p(D|\lambda)}, \quad p_2(Q) = \frac{p(Q, D|\lambda')}{p(D|\lambda')}.$$

- Тогда p_1 и p_2 — распределения, и расстояние Kullback–Leibler:

$$\begin{aligned} 0 \leq D_{LK}(\lambda, \lambda') &= \sum_Q \frac{p(Q, D|\lambda)}{p(D|\lambda)} \log \frac{p(Q, D|\lambda)p(D|\lambda')}{p(Q, D|\lambda')p(D|\lambda)} = \\ &= \log \frac{p(D|\lambda')}{p(D|\lambda)} + \sum_Q \frac{p(Q, D|\lambda)}{p(D|\lambda)} \log \frac{p(Q, D|\lambda)}{p(Q, D|\lambda')}. \end{aligned}$$

- Введём вспомогательную функцию

$$Q(\lambda, \lambda') = \sum_Q p(Q|D, \lambda) \log p(Q|D, \lambda').$$

- Тогда из неравенства следует, что

$$\frac{Q(\lambda, \lambda') - Q(\lambda, \lambda)}{p(D|\lambda)} \leq \log \frac{p(D|\lambda')}{p(D|\lambda)}.$$

- Т.е., если $Q(\lambda, \lambda') > Q(\lambda, \lambda)$, то $p(D|\lambda') > p(D|\lambda)$.
- Т.е., если мы максимизируем $Q(\lambda, \lambda')$ по λ' , мы тем самым будем двигаться в нужную сторону.

- Нужно максимизировать $Q(\lambda, \lambda')$. Перепишем:

$$\begin{aligned} Q(\lambda, \lambda') &= \sum_Q p(Q|D, \lambda) \log p(Q|D, \lambda') = \\ &= \sum_Q p(Q|D, \lambda) \log \pi_{q_1} \prod_t a_{q_{t-1}q_t} b_{q_t}(d_t) = \\ &= \sum_Q p(Q|D, \lambda) \log \pi_{q_1} + \sum_Q p(Q|D, \lambda) \sum_t \log a_{q_{t-1}q_t} b_{q_t}(d_t). \end{aligned}$$

- Последнее выражение легко дифференцировать по a_{ij} , $b_i(k)$ и π_i , добавлять соответствующие множители Лагранжа и решать. Получится именно пересчёт по алгоритму Баума–Велха (проверьте!).

СПЕЦИАЛЬНЫЕ ВИДЫ МАРКОВСКИХ МОДЕЛЕЙ

- У нас были дискретные наблюдаемые с вероятностями $B = (b_j(k))$.
- Но в реальной жизни всё сложнее: зачастую мы наблюдаем непрерывные сигналы, а не дискретные величины, и дискретизовать их или плохо, или неудобно.
- При этом саму цепь можно оставить дискретной, т.е. перейти к непрерывным $b_j(D)$.

СПЕЦИАЛЬНЫЙ ВИД ПЛОТНОСТИ

- Не для всех плотностей найдены алгоритмы пересчёта (обобщения алгоритма Баума–Велха).
- Наиболее общий результат верен, когда $b_j(D)$ можно представить в виде

$$b_j(D) = \sum_{m=1}^M c_{jm} \mathcal{P}(D, \mu_{jm}, \sigma_{jm}),$$

где c_{jm} — коэффициенты смеси ($\sum_m c_{jm} = 1$), а \mathcal{P} — выпуклое распределение со средним μ и вариацией σ (гауссиан подойдёт).

- К счастью, такой конструкцией можно приблизить любое непрерывное распределение, поэтому это можно широко применять.

- $\gamma_t(j, m)$ — вероятность быть в состоянии j во время t , причём за D отвечает m -й компонент смеси.
- Формально говоря,

$$\gamma_t(j, m) = \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^N \alpha_t(j)\beta_t(j)} \right] \left[\frac{c_{jm}\mathcal{P}(d_t, \mu_{jm}, \sigma_{jm})}{\sum_{m=1}^M c_{jm}\mathcal{P}(d_t, \mu_{jm}, \sigma_{jm})} \right] \cdot$$

- Если $M = 1$, то это уже известные нам $\gamma_t(j)$.

- Нужно научиться пересчитывать $b_j(D)$, т.е. пересчитывать c_{jm} , μ_{jm} и σ_{jm} .
- Это делается так:

$$\bar{c}_{jm} = \frac{\sum_{t=1}^T \gamma_t(j, m)}{\sum_{t=1}^T \sum_{m=1}^M \gamma_t(j, m)},$$

$$\bar{\mu}_{jm} = \frac{\sum_{t=1}^T \gamma_t(j, m) \cdot d_t}{\sum_{t=1}^T \gamma_t(j, m)},$$

$$\bar{\sigma}_{jm} = \frac{\sum_{t=1}^T \gamma_t(j, m) \cdot (d_t - \mu_{jm})(d_t - \mu_{jm})^t}{\sum_{t=1}^T \gamma_t(j, m)}.$$

- Как моделировать продолжительность нахождения в том или ином состоянии?
- В дискретном случае вероятность пробыть в состоянии i d шагов:

$$p_i(d) = a_{ii}^{d-1}(1 - a_{ii}).$$

- Однако для большинства физических сигналов такое экспоненциальное распределение не соответствует действительности. Мы бы хотели явно задавать плотность пребывания в данном состоянии.
- Т.е. вместо коэффициентов перехода в себя a_{ii} — явное задание распределения $p_i(d)$.

- Введём переменные

$$\alpha_t(i) = p(d_1 \dots d_t, x_i \text{ заканчивается во время } t|\lambda).$$

- Всего за первые t шагов посещено r состояний $q_1 \dots q_r$, и мы там оставались d_1, \dots, d_r . Т.е. ограничения:

$$q_r = x_i, \quad \sum_{s=1}^r d_s = t.$$

- Тогда получается

$$\alpha_t(i) = \sum_q \sum_d \pi_{q_1} p_{q_1}(d_1) p(d_1 d_2 \dots d_{d_1} | q_1) \\ a_{q_1 q_2} p_{q_2}(d_2) p(d_{d_1+1} \dots d_{d_1+d_2} | q_2) \dots \\ \dots a_{q_{r-1} q_r} p_{q_r}(d_r) p(d_{d_1+\dots+d_{r-1}+1} \dots d_t | q_r).$$

- По индукции

$$\alpha_t(j) = \sum_{i=1}^n \sum_{d=1}^D \alpha_{t-d}(j) a_{ij} p_j(d) \prod_{s=t-d+1}^t b_j(d_s),$$

где D — максимальная остановка в любом из состояний.

- Тогда, как и раньше,

$$p(d|\lambda) = \sum_{i=1}^n \alpha_T(i).$$

- Для пересчёта потребуются ещё три переменные:

$$\alpha_t^*(i) = p(d_1 \dots d_t, x_i \text{ начинается во время } t + 1 | \lambda),$$

$$\beta_t(i) = p(d_{t+1} \dots d_T | x_i \text{ заканчивается во время } t, \lambda),$$

$$\beta_t^*(i) = p(d_{t+1} \dots d_T | x_i \text{ начинается во время } t + 1, \lambda).$$

- Соотношения между ними:

$$\alpha_t^*(j) = \sum_{i=1}^n \alpha_t(i) a_{ij},$$

$$\alpha_t(i) = \sum_{d=1}^D \alpha_{t-d}^*(i) p_i(d) \prod_{s=t-d+1}^t b_i(d_s),$$

$$\beta_t(i) = \sum_{j=1}^n a_{ij} \beta_t^*(j),$$

$$\beta_t^*(i) = \sum_{d=1}^D \beta_{t+d}(i) p_i(d) \prod_{s=t+1}^{t+d} b_i(d_s).$$

- Приведём формулы пересчёта.
- π_i — просто вероятность того, что x_i был первым состоянием:

$$\hat{\pi}_i = \frac{\pi_i \beta_0^*(i)}{p(d|\lambda)}.$$

- a_{ij} — та же формула, что обычно, только вместе с α есть ещё и β , которая говорит, что новое состояние начинается на следующем шаге:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^T \alpha_t(i) a_{ij} \beta_t^*(j)}{\sum_{k=1}^n \sum_{t=1}^T \alpha_t(i) a_{ik} \beta_t^*(k)}.$$

- $b_i(k)$ — отношение ожидания количества событий $d_t = v_k$ в состоянии x_i к ожиданию количества любого v_j в состоянии x_i :

$$\hat{b}_i(k) = \frac{\sum_{t=1, d_t=v_k}^T \left(\sum_{\tau < t} \alpha_\tau^*(i) \beta_\tau^*(i) - \sum_{\tau < t} \alpha_\tau(i) \beta_\tau(i) \right)}{\sum_{k=1}^m \sum_{t=1, d_t=v_k}^T \left(\sum_{\tau < t} \alpha_\tau^*(i) \beta_\tau^*(i) - \sum_{\tau < t} \alpha_\tau(i) \beta_\tau(i) \right)}.$$

- $p_i(d)$ — отношение ожидания количества раз, которые x_i случилось с продолжительностью d , к количеству раз, которые x_i вообще случилось:

$$\hat{p}_i(d) = \frac{\sum_{t=1}^T \alpha_t^*(i) p_i(d) \beta_{t+d}(i) \prod_{s=t+1}^{t+d} b_i(d_s)}{\sum_{d=1}^D \sum_{t=1}^T \alpha_t^*(i) p_i(d) \beta_{t+d}(i) \prod_{s=t+1}^{t+d} b_i(d_s)}.$$

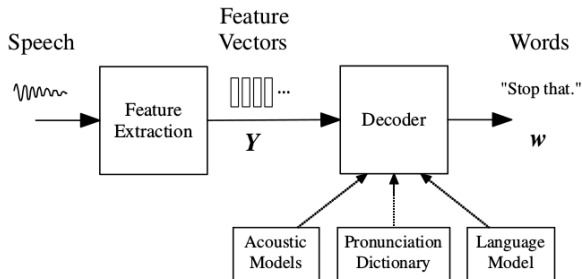
- Такой подход очень полезен, когда $p_i(d)$ далеко от экспоненциального.
- Однако он сильно увеличивает вычислительную сложность (в D^2 раз).
- И, кроме того, становится гораздо больше параметров, т.е. нужно, вообще говоря, больше данных, чтобы эти параметры надёжно оценить.

- Чтобы уменьшить количество параметров, можно иногда считать, что $p_i(d)$ — классическое распределение с не слишком большим количеством параметров.
- Например, $p_i(d)$ может быть равномерным, или нормальным ($p_i(d) = \mathcal{N}(d, \mu_i, \sigma_i^2)$), или гамма-распределением:

$$p_i(d) = \frac{\eta_i^{\nu_i} d^{\nu_i-1} e^{-\eta_i d}}{\Gamma(\nu_i)}.$$

НММ ДЛЯ РАСПОЗНАВАНИЯ РЕЧИ

- НММ – классический подход к распознаванию речи.
- Сейчас, правда, они уже в основном заменены глубокими нейронными сетями; но всё равно полезно проследить, как их можно применить.



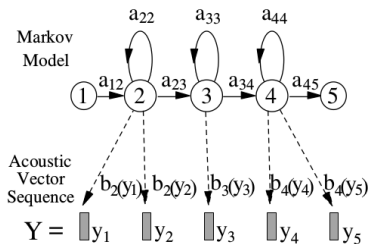
- Признаки как-то выделились, а потом слово w делится на фонемы $\mathbf{q} = q_1 \dots q_{|w|}$, и правдоподобие наблюдаемых признаков

$$p(\mathbf{y} | \mathbf{w}) = \sum_{\mathbf{q}} p(\mathbf{y} | \mathbf{q})p(\mathbf{q} | \mathbf{w}),$$

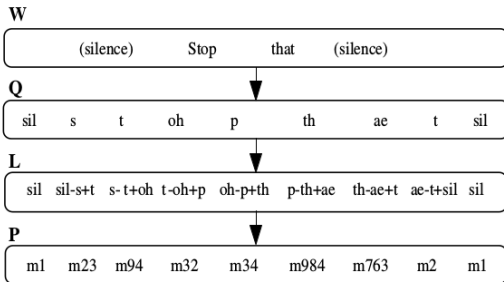
сумма по возможным произношениям (маленькая сумма), а \mathbf{w} – это все слова $w_1 \dots w_L$:

$$p(\mathbf{q} | \mathbf{w}) = \prod_{l=1}^L p(\mathbf{q}^{(w_l)} | w_l).$$

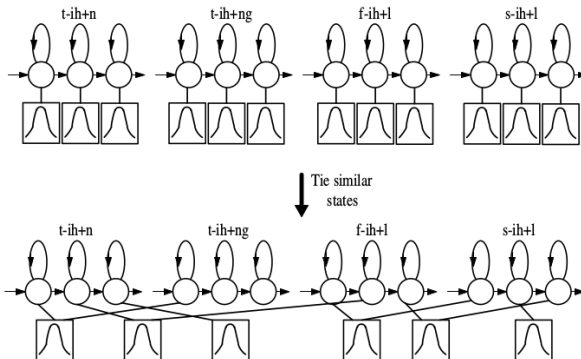
- Каждая фонема – это НММ с непрерывными наблюдаемыми $b_j(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mu^{(j)}, \Sigma^{(j)})$.
- На этом месте уже можно обучать просто всё сразу.



- Однако фонемы очень по-разному звучат в зависимости от контекста.
- Можно перейти к трифонам.

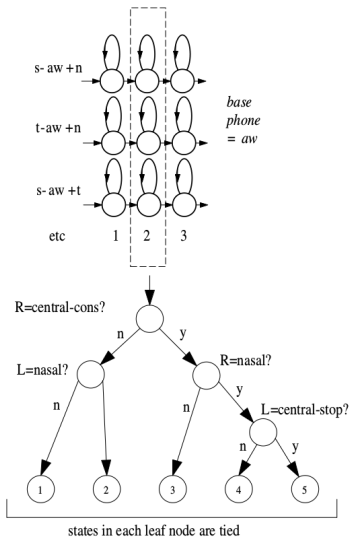


- Но их будет целых N^3 , и лучше объединить похожие и связать их параметры:



АКУСТИЧЕСКАЯ МОДЕЛЬ

- Это можно сделать просто силой мысли:



- Но это только начало. Ещё нужна *языковая модель* – мы о многом просто догадываемся.
- То есть нужно априорное распределение

$$p(\mathbf{w}) = \prod_{l=1}^L p(w_l | w_1 \dots w_{l-1}).$$

- Классический подход – n -граммы для $n = 2..4$:

$$p(\mathbf{w}) = \prod_{l=1}^L p(w_l | w_{l-1} \dots w_{l-n+1}).$$

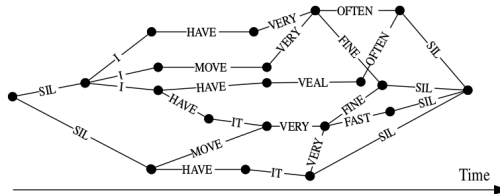
- Качество языковых моделей сравнивают в терминах их *перплексии* (perplexity)

$$H = - \lim_{L \rightarrow \infty} \frac{1}{K} \log p(w_1, \dots, w_L).$$

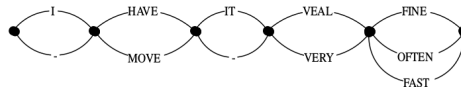
- О современных языковых моделях мы обязательно поговорим потом...
- А пока декодер идёт и алгоритмом Витерби всё решает.
- Но что именно решает?

- Возможности удобно представлять как *решётку слов* (word lattice) или *confusion network*.

(a) Word Lattice

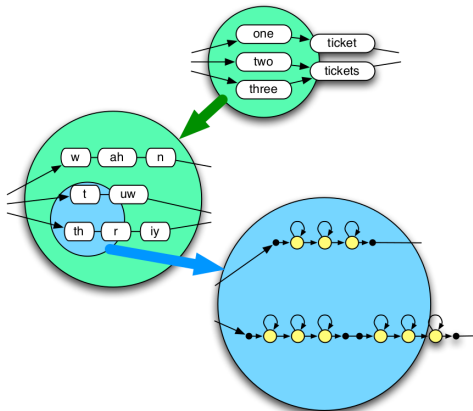


(b) Confusion Network

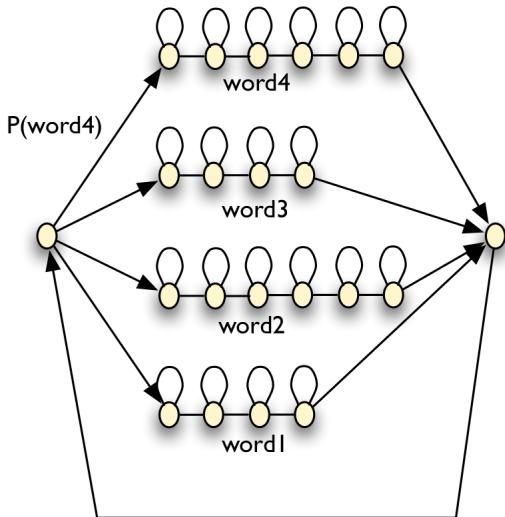


РЕЗУЛЬТАТЫ

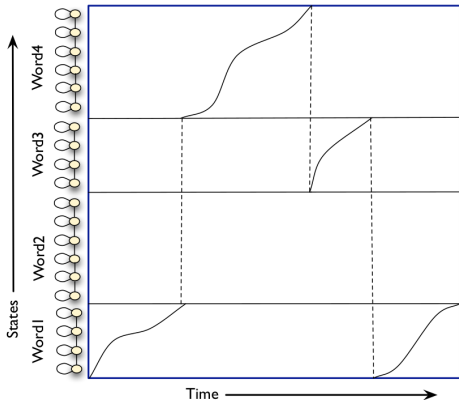
- Сеть слов превращается в сеть фонем, потом в сеть состояний HMM.



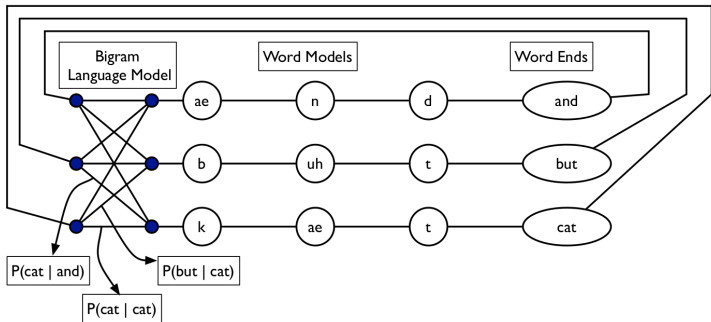
- И можно распознавать связанные друг с другом слова тоже алгоритмом Витерби.



- Всё это накладывается на собственно аудиозапись, получается последовательность во времени.

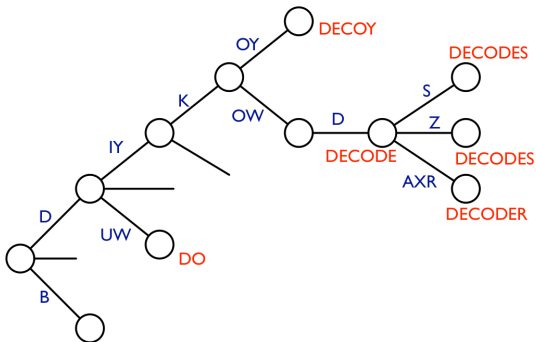


- А языковая модель – это просто дополнительные множители (слагаемые в \log), априорные вероятности.



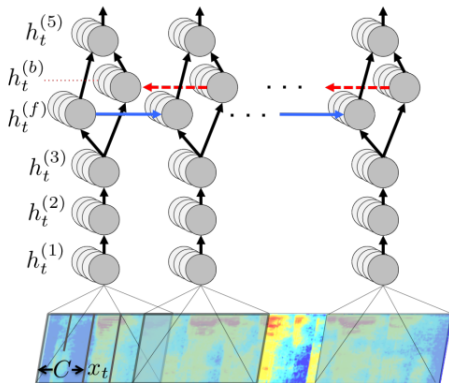
- Декодирование по всей сети всего языка нереально.
- Обычно декодер генерирует и поддерживает только лучшие гипотезы; это называется *beam search*.
- Т.е. мы после каждого шага (обычно на границах слов) делаем pruning и либо выкидываем гипотезы, которые сильно хуже текущего лучшего варианта, либо просто поддерживаем N лучших (типа 1000).

- А НММ для отдельных слов (нам же нужно по НММ на каждое слово) тоже можно организовать в дерево (префиксное).



END-TO-END

- Впрочем, сейчас уже всё по-другому.
- End-to-end speech recognition.
- Но об этом – потом.



Спасибо за внимание!