Лекция 12. Вариационный подход для оценки обоснованности

Д. П. Ветров[1] Д. А. Кропотов[2]

[1]МГУ, ВМиК, каф. ММП

[2]ВЦ РАН

Спецкурс «Байесовские методы машинного обучения»

План лекции

План лекции

Дивергенция Кульбака-Лейблера

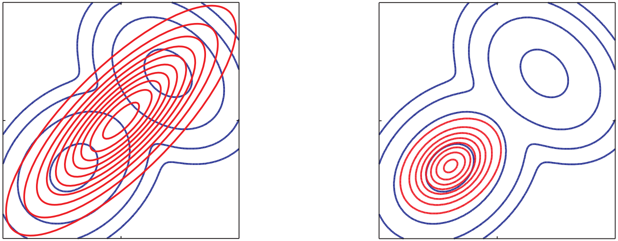
- Существует множество способов определить близость между вероятностными распределениями
- Рассмотрим распределения $p(\boldsymbol{x})$ и $q(\boldsymbol{x})$. Дивергенцией Кульбака-Лейблера называется величина

$$KL(q||p) = -\int q(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}$$

- $KL(p||q) \geq 0$ для любых двух распределений
- Дивергенция равна нулю тогда и только тогда, когда $q(\boldsymbol{x}) = p(\boldsymbol{x})$
- Антисимметричность: $KL(p||q) \neq KL(q||p)$
- Минимизация дивергенции Кульбака-Лейблера часто используется для приближения сложного распределения $p(\boldsymbol{x})$ более простым распределением $q(\boldsymbol{x})$

Геометрический смысл

Пусть $p(\boldsymbol{x})$ — «синее» распределение, а $q(\boldsymbol{x})$ — «красное». Слева показан результат минимизации $KL(p||q)$ по $q(\boldsymbol{x})$, а справа — результат минимизации $KL(q||p)$ по $q(\boldsymbol{x})$



Гамма-распределение

- Гамма-распределение имеет плотность

$$\mathcal{G}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda), \quad a,b > 0$$

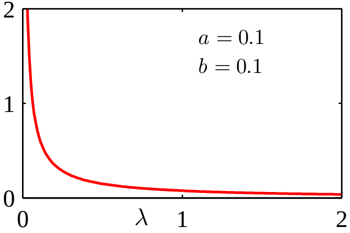
- Характеристики гамма-распределения

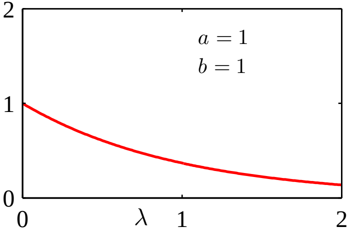
$$\mathbb{E}\lambda = \frac{a}{b}, \qquad \mathbb{D}\lambda = \frac{a}{b^2}$$

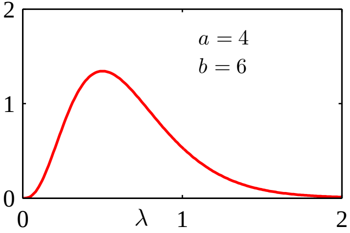
- Гамма-распределение является сопряженным для обратной дисперсии (точности) нормального распределения $\lambda = \sigma^{-2}$, т.к.

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

График гамма-распределения







План лекции

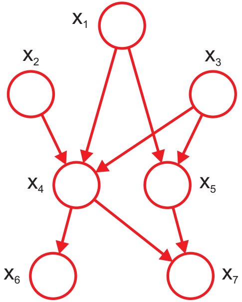
Байесовская сеть

- Совместное распределение переменных задается с помощью ориентированного графа
- У графа не может быть ориентированных циклов
- Каждая вершина графа соответствует случайной величине
- Каждая стрелка соответствует причинно-следственной взаимосвязи между переменными
- В общем случае, совместное распределение для графа с n вершинами

$$p(\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i|\mathrm{pa}_i),$$

где pa_i — множество вершин-родителей x_i

Пример байесовской сети



Совместное распределение системы переменных задается выражением

$$p(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = \\ p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5).$$

Вывод в байесовской сети

- Обозначим совокупность переменных графической модели через (X, T, Z). Здесь X — множество наблюдаемых переменных, T — множество ненаблюдаемых переменных, подлежащих оценке, а Z — множество ненаблюдаемых переменных, конкретные значения которых нас не интересуют.
- Тогда задача вывода в графической модели формулируется как поиск условного распределения

$$p(T|X) = \int p(T, Z|X)dZ = \frac{\int P(T, Z, X)dZ}{p(X)} = \frac{\int p(T, Z, X)dZ}{\int p(T, Z, X)dTdZ}$$

- При работе с графическими моделями широко используются правило суммы и произведения:

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

Сравнение вероятностных моделей

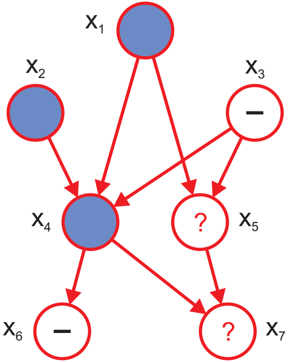
Другой задачей в графических моделях является вычисление обоснованности (правдоподобия):

$$p(X) = \int p(X, Z, T) dZdT$$

Правдоподобие показывает меру адекватности модели наблюдаемым данным. Данная величина может быть использована для сравнения различных моделей (пример — поиск коэффициента регуляризации α в методе релевантных векторов)

Вычисление условных распределений I

Пусть нам необходимо найти распределение (x_5, x_7) при заданных значениях x_1, x_2, x_4 и неизвестных x_3, x_6



Вычисление условных распределений II

- По правилу произведения

$$p(x_5, x_7|x_1, x_2, x_4) = \frac{p(x_1, x_2, x_4, x_5, x_7)}{p(x_1, x_2, x_4)}$$

- Расписываем знаменатель

$$p(x_1, x_2, x_4) = p(x_1)p(x_2)p(x_4|x_1, x_2) = \{Sum\ rule\}$$

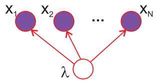
$$p(x_1)p(x_2)\int p(x_4|x_1, x_2, x_3)p(x_3)dx_3$$

- Аналогично числитель

$$p(x_1, x_2, x_4, x_5, x_7) = p(x_1)p(x_2)p(x_4|x_1, x_2)p(x_5|x_1)p(x_7|x_5, x_4) = p(x_1)\times$$
$$p(x_2)\left(\int p(x_4|x_1, x_2, x_3)p(x_3)dx_3\right)\left(\int p(x_5|x_1, x_3)p(x_3)dx_3\right)p(x_7|x_5, x_4)$$

- Для взятия возникающих интегралов обычно пользуются приближением Лапласа, методами Монте Карло или вариационным подходом.
- Таким образом, условное распределение выражено через известные атомарные распределения вида $p(x_i|pa_i)$

Пример байесовского оценивания



Рассмотрим следующую вероятностную модель:

$$p(x|\lambda) = \mathcal{N}(x|0,\lambda) = \sqrt{\frac{\lambda}{2\pi}}\exp\left(-\frac{\lambda}{2}x^2\right)$$

$$p(\lambda) = \mathcal{G}(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda)$$

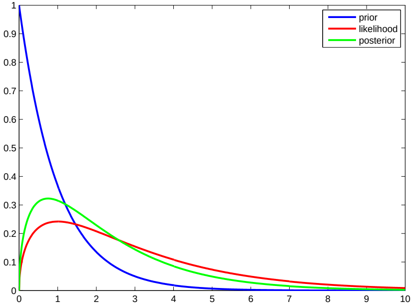
Вывод в модели:

$$p(\lambda|X) = \frac{p(X,\lambda)}{p(X)} = \frac{p(X|\lambda)p(\lambda)}{p(X)}$$

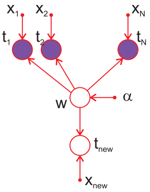
Продолжение примера

$$p(\lambda|X) = \frac{p(X|\lambda)p(\lambda)}{p(X)} = \frac{1}{p(X)}\left(\frac{\lambda}{2\pi}\right)^{N/2}\exp\left(-\frac{\lambda}{2}\sum_{i=1}^{N}x_i^2\right)\times$$

$$\frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-b\lambda) = \mathcal{G}\left(\lambda\,\middle|\,a+\frac{N}{2}, b+\frac{1}{2}\sum_{i=1}^{N}x_i^2\right)$$



Метод релевантных векторов



Вероятностная модель:

$$p(t_i|\boldsymbol{x}_i) = \mathcal{N}(t_i|y(\boldsymbol{w}, \boldsymbol{x}_i), \sigma^2)$$
$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}I)$$

Вывод в модели:

$$p(t_{new}|T, X, \boldsymbol{x}_{new}, \alpha, \sigma) = \int p(t_{new}|\boldsymbol{w}, \sigma^2)p(\boldsymbol{w}|T, X, \alpha)d\boldsymbol{w}$$

Обучение в модели:

$$p(T|\alpha, X) = \int p(T|X, \boldsymbol{w})p(\boldsymbol{w}|\alpha) \to \max_{\alpha}$$

План лекции

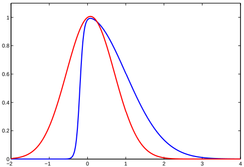
Приближение Лапласа

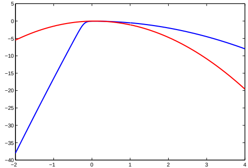
- Рассмотрим функцию $p(z) = \exp\left(-\frac{z^2}{2}\right)\frac{1}{1+\exp(-20z-4)}$.
- Разложим логарифм функции в ряд Тейлора в точке максимума:

$$z_0 = \arg\max_z f(z),\ \log f(z) \simeq \log f(z_0) + \frac{H}{2}(z-z_0)^2,\ H = \left.\frac{d^2 \log f}{dz^2}\right|_{z=z_0}$$

- Тогда функцию $f(z)$ можно приблизить следующим образом:

$$f(z) \simeq f(z_0)\exp\left(\frac{H}{2}(z-z_0)^2\right)$$





Недостатки приближения Лапласа

- Метод Лапласа хорошо приближает распределение гауссианой в точке максимума, но плохо делает приближение в целом, если распределение сильно отличается от гауссианы
- В частности, математические ожидания и дисперсии распределения и его приближения Лапласа могут сильно отличаться
- Это приводит к сильным смещениям оценки обоснованности

Приближение апостериорного распределения

- Вероятностная модель обычно позволяет в явном виде задать совместное распределение $p(X, Z)$. Целью задачи является нахождение (или приближение) обоснованности выбранной модели $p(X) = \int P(X, Z)dZ$ и апостериорного распределения

$$p(Z|X) = \frac{p(X, Z)}{p(X)}$$

- На практике прямое интегрирование выражения $p(X, Z)$ обычно невозможно, поэтому ограничиваются приближением распределения $p(Z|X)$ с помощью некоторого распределения $q(Z)$

Разложение обоснованности

- Справедливо следующее преобразование

$$\log p(X) = \log p(X) \int q(Z)dZ = \int \log p(X) q(Z) dZ =$$
$$\int \log \frac{p(X,Z)}{p(Z|X)} q(Z)dZ = \int \log \frac{p(X,Z)q(Z)}{q(Z)p(Z|X)} q(Z)dZ =$$
$$\int \log \frac{p(X,Z)}{q(Z)} q(Z)dZ - \int \log \frac{p(Z|X)}{q(Z)} q(Z)dZ = \mathcal{L}(q) + KL(q||p)$$

- Величина $\mathcal{L}(q)$ представляет собой нижнюю границу логарифма обоснованности
- Так как $\log p(X)$ не зависит от $q(Z)$, максимизация $\mathcal{L}(q)$ эквивалентна минимизации дивергенции Кульбака-Лейблера $KL(q||p)$ между $q(Z)$ и апостериорным распределением $p(Z|X)$!

Поиск приближения

- Поиск приближения $q(Z)$ можно осуществлять путем минимизации $KL(q(Z)||p(Z|X))$. Эквивалентно, эта задача может быть решена путем максимизации $L(q)$ или, что тоже самое

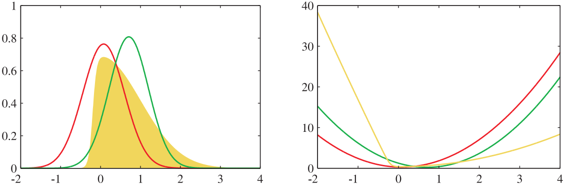
$$-L(q) = KL(q(Z)||p(X,Z)) \to \min_{q(Z)}$$

  Эта задача, вообще говоря, проще, т.к. $p(X,Z)$ — известно в отличие от $p(Z|X)$.
- Для оценки обоснованности $p(X)$ можно также решать задачу минимизации $KL(q(Z)||p(X,Z))$, т.к. величина $L(q)$ является нижней границей обоснованности:

$$p(X) \geq L(q)$$

- Решать задачу минимизации дивергенции KL можно в рамках параметрического семейства. Другой подход состоит в рассмотрении семейства факторизованных распределений

Пример



— исходное распределение
— приближение Лапласа
— вариационное приближение в семействе нормальных распределений

Факторизация $q(Z)$

Разобьем множество переменных Z на непересекающиеся группы $Z = (Z_1, \dots, Z_k)$, $Z_i \cap Z_j = \emptyset$, $i \neq j$.

Рассмотрим в качестве семейства для $q(Z)$ факторизованное семейство:

$$q(Z) = \prod_{i=1}^{k} q_i(Z_i)$$

Тогда можно показать, что задача минимизации дивергенции в таком семействе может быть решена напрямую без использования параметризации.

Факторизованное приближение

- Рассмотрим задачу минимизации $KL(q(Z)||p(Z))$ по одной компоненте q_j.
- Подставим $q(Z) = \prod_{i=1}^k q_i(Z_i) = \prod_{i=1}^k q_i$ в выражение для $KL(q||p)$

$$KL(q(Z)||p(Z)) = -\int \prod_i q_i \left(\log p(Z) - \sum_i \log q_i \right) dZ =$$

$$-\int q_j \left(\int \log p(Z) \prod_{i \neq j} q_i dZ_i \right) dZ_j + \int q_j \log q_j dZ_j + C$$

- Обозначим $\log \tilde{p}(Z_j) = \mathbb{E}_{i \neq j} \log p(Z) + C = \int \log p(Z) \prod_{i \neq j} q_i dZ_i + C$. Тогда

$$KL(q||p) = -\int q_j \log \frac{\tilde{p}(Z_j)}{q_j} dZ_j + C = KL(q_j||\tilde{p}) + C$$

Основной результат

$$KL(q||p) = KL(q||\tilde{p}) + C \to \min_{q_j}$$

- Оптимальное распределение $q_j^*(Z_j) = \tilde{p}(Z_j)$, т.е.

$$\log q_j^*(Z_j) = \mathbb{E}_{i\neq j} \log p(Z) + C$$

$$q_j^*(Z_j) = \frac{\exp(\mathbb{E}_{i\neq j} \log p(Z))}{\int \exp(\mathbb{E}_{i\neq j} \log p(Z)) dZ_j}$$

- Заметим, что нам не пришлось делать каких-либо предположений о функциональной форме распределения $q_j(Z_j)$
- Выражение для оптимального $q_j^*(Z_j)$ зависит от остальных $q_i(Z_i)$, поэтому необходима итерационная оптимизация
- Часто константа C не требует оценки, а определяется из вида распределения

Вероятностная модель линейной регрессии

- Рассмотрим стандартную задачу восстановления регрессии $(X, \boldsymbol{t})$ — обучающая выборка, $t \in \mathbb{R}$. Регрессия имеет вид $y(\boldsymbol{x}) = \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$
- Определим следующую вероятностную модель $p(\boldsymbol{t}, \boldsymbol{w}, \alpha) = p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}|\alpha)p(\alpha)$, где

$$p(\boldsymbol{t}|\boldsymbol{w}) = \prod_{i=1}^{n} \mathcal{N}(t_i|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i), \beta^{-1}) \qquad p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}I)$$

$$p(\alpha) = \mathcal{G}(\alpha|a_0, b_0)$$

- В данной модели роль наблюдаемых переменных играет $\boldsymbol{t}$, а в роли Z выступают $\boldsymbol{w}$ и α
- Для простоты предположим, что значение интенсивности белого шума β известно

Вариационный вывод для α

- Будем искать приближение распределения $p(\boldsymbol{w},\alpha|\boldsymbol{t})$ в виде
$$q(\boldsymbol{w},\alpha)=q(\boldsymbol{w})q(\alpha)$$
- Используя основной результат для $q(\alpha)$ получаем
$$\log q^*(\alpha)=\mathbb{E}_{\boldsymbol{w}}\log p(\boldsymbol{t},\boldsymbol{w},\alpha)=\mathbb{E}_{\boldsymbol{w}}\left(\log p(\boldsymbol{w}|\alpha)p(\alpha)\right)+C=$$
$$\mathbb{E}_{\boldsymbol{w}}\log p(\boldsymbol{w}|\alpha)+\log p(\alpha)+C=$$
$$\frac{m}{2}\log\alpha-\frac{\alpha}{2}\mathbb{E}\boldsymbol{w}^T\boldsymbol{w}+(a_0-1)\log\alpha-b_0\alpha+C_1$$
- Но это в точности логарифм гамма-распределения с параметрами a_n и b_n, т.е. $\alpha\sim\mathcal{G}(\alpha|a_n,b_n)$, причем
$$a_n=a_0+\frac{m}{2},\qquad b_n=b_0+\frac{1}{2}\mathbb{E}\boldsymbol{w}^T\boldsymbol{w}$$

Вариационный вывод для $\boldsymbol{w}$

- Проделаем аналогичную операцию для $q(\boldsymbol{w})$

$$\log q^*(\boldsymbol{w}) = \mathbb{E}_\alpha \log p(\boldsymbol{t}, \boldsymbol{w}, \alpha) = \mathbb{E}_\alpha \log\left(p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}|\alpha)p(\alpha)\right) =$$
$$\log p(\boldsymbol{t}|\boldsymbol{w}) + \mathbb{E}_\alpha \log p(\boldsymbol{w}|\alpha) + C =$$
$$-\frac{\beta}{2}\sum_{i=1}^{n}(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) - t_i)^2 - \frac{1}{2}\mathbb{E}\alpha \cdot \boldsymbol{w}^T\boldsymbol{w} + C_1 =$$
$$-\frac{1}{2}\boldsymbol{w}^T(\mathbb{E}\alpha I + \beta\Phi^T\Phi)\boldsymbol{w} + \beta\boldsymbol{w}^T\Phi^T\boldsymbol{t} + C_2,$$

  где $\Phi = (\phi(\boldsymbol{x}_1), \dots, \phi(\boldsymbol{x}_n))$
- Последовательно проведено отбрасывание слагаемых, не зависящих от $\boldsymbol{w}$, раскрытие скобок и приведение подобных слагаемых
- Выделяя полный квадрат, получаем, что $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_n, S_n)$, где

$$\boldsymbol{\mu}_n = \beta S_n\Phi^T\boldsymbol{t}, \qquad S_n = (\mathbb{E}\alpha I + \beta\Phi^T\Phi)^{-1}$$

Итерационные формулы

- Окончательные формулы: $q^*(\alpha) = \mathcal{G}(\alpha|a_n, b_n)$, $q^*(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_n, S_n)$, т.е.

$$\mathbb{E}\alpha = \frac{a_n}{b_n}$$

$$\mathbb{E}\boldsymbol{w}^T\boldsymbol{w} = \mathrm{tr}(\mathbb{E}\boldsymbol{w}\boldsymbol{w}^T) = \mathrm{tr}(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^T + S_n) = \boldsymbol{\mu}_n^T\boldsymbol{\mu}_n + \mathrm{tr}\, S_n$$

- Параметры распределений определяются по итерационным формулам

$$a_n = a_0 + \frac{m}{2}$$

$$b_n = b_0 + \mathbb{E}\boldsymbol{w}^T\boldsymbol{w} = b_0 + \boldsymbol{\mu}_n^T\boldsymbol{\mu}_n + \mathrm{tr}\, S_n$$

$$\boldsymbol{\mu}_n = \beta S_n \Phi^T \boldsymbol{t}$$

$$S_n = \left(\mathbb{E}\alpha I + \beta\Phi^T\Phi\right)^{-1} = \left(\frac{a_n}{b_n} I + \beta\Phi^T\Phi\right)^{-1}$$

Заключительные замечания

- Отметим, что никаких ограничений на форму апостериорных распределений не вводилось, а единственным приближением было предположение о факторизации
- Вариационный метод позволяет получать приближение обоснованности, нижней оценкой которой является выражение

$$\mathcal{L}(q) = \int q(\boldsymbol{w}, \alpha) \log \frac{p(\boldsymbol{t}, \boldsymbol{w}, \alpha)}{q(\boldsymbol{w}, \alpha)} d\boldsymbol{w} d\alpha =$$

$$\mathbb{E} \log p(\boldsymbol{t}, \boldsymbol{w}, \alpha) - \mathbb{E} \log q(\boldsymbol{w}, \alpha) =$$

$$\mathbb{E}_{\boldsymbol{w}} \log p(\boldsymbol{t}|\boldsymbol{w}) + \mathbb{E}_{\boldsymbol{w},\alpha} \log p(\boldsymbol{w}|\alpha) + \mathbb{E}_{\alpha} \log p(\alpha)$$

$$- \mathbb{E}_{\boldsymbol{w}} \log q(\boldsymbol{w}) - \mathbb{E}_{\alpha} q(\alpha)$$

- Все эти выражения выписываются в явном виде (Упр.)

Вариационный выбор модели

На рисунке изображена зависимость $\mathcal{L}(q)$ от степени полинома для полиномиальной регрессии, построенной по зашумленной выборке, полученной с помощью кубического многочлена