

Прикладная статистика в машинном обучении

Лекция 10

Метод релевантных векторов

И. К. Козлов
(*Мехмат МГУ*)

2022

Настало время вспомнить **Лекцию 2**
про вероятностный подход к регрессии.



Рис.: Hello darkness, my old friend

Иногда они возвращаются

Нужно выполнить данное обещание:

Обсудим во 2ой части курса.

Если ввести правильное априорное распределение на веса $p(\beta)$
и применить теорему Байеса

$$p(\beta \mid \text{Data}) = \frac{p(\text{Data} \mid \beta)p(\beta)}{p(\text{Data})}$$

получатся формулы для регуляризации.

Литература

Байесовская линейная регрессия и Метод релевантных векторов (RVM) обсуждаются в Главах 3.4 и 7.2

 Bishop C.M.
Pattern Recognition and Machine Learning.

Будет сложно, надеюсь, Вам понравится

Эта и последующие лекции — самые сложные.

К тому же в этой лекции — очень много формул.

Не расстраивайтесь, если в них не удастся быстро разобраться.



Disclaimer

Чтобы не потонуть в формулах.

Теорема

Каждый раздел подводит к утверждению, выделенному такой рамочкой.

Можно вначале осознать утверждение, потом — по возможности доказательство.

Как лучше читать

Формул много, но они однотипные.

- Все распределения в этой лекции — нормальные.
- Большинство вычислений — перемножения двух нормальных распределений

$$\mathcal{N}(\mu_1, \sigma_1^2) \mathcal{N}(\mu_2, \sigma_2^2).$$

Линейная регрессия по-статистически

1 Линейная регрессия по-статистически

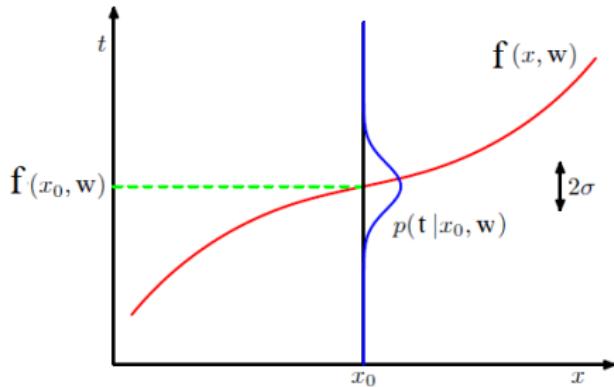
2 Линейная регрессия по-байесовски

- МАР-оценка
- Аналитический байесовский вывод

3 Метод релевантных векторов

- Вариационная нижняя оценка
- Алгоритм

Линейная регрессия



Знакомые нам формулы:

- Мы наблюдаем зашумлённый таргет:

$$t_i = w^T x + \varepsilon_i,$$

- где ε_i — гауссовский шум

$$\varepsilon_i | X \sim \mathcal{N}(0, \sigma^2).$$

When in Rome...



Q: Как записать “в байесовском виде” эту модель?

$$t_i = w^T x + \varepsilon_i, \quad \varepsilon_i | X \sim \mathcal{N}(0, \sigma^2).$$

Нужно написать правильную условную вероятность.

(Какое распределение? Что фиксировано?)

Условная вероятность

“При фиксированных x и w таргет t нормально распределён”:

$$p(t | x, w) = \mathcal{N}(t | w^T x, \beta^{-1}).$$

Замечание. Дисперсия обозначена за β^{-1} для удобства дальнейших вычислений.

Условная вероятность

“При фиксированных x и w таргет t нормально распределён”:

$$p(t | x, w) = \mathcal{N}(t | w^T x, \beta^{-1}).$$

Замечание. Дисперсия обозначена за β^{-1} для удобства дальнейших вычислений.

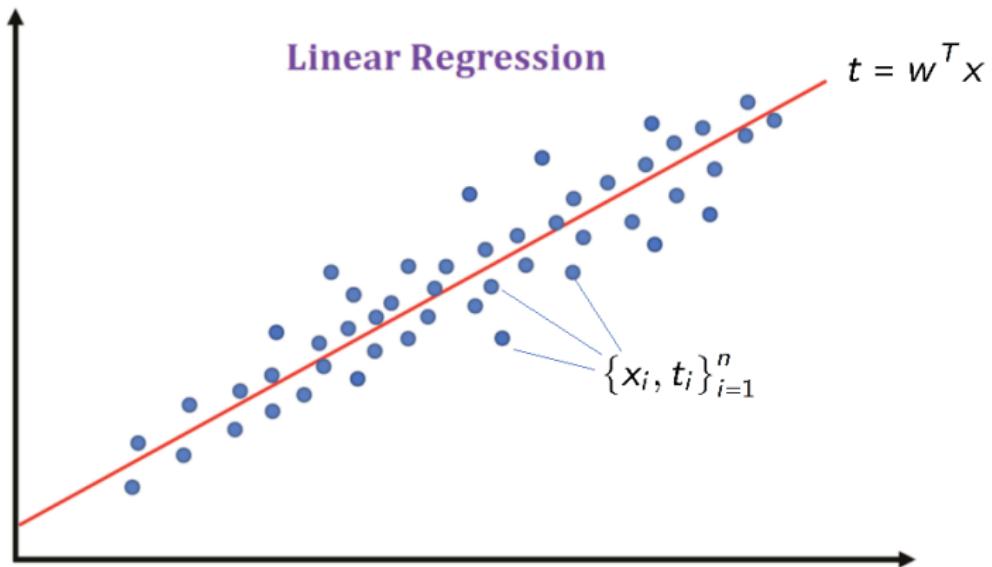
Q: Что нам дано и что мы хотим найти?

Выборка и параметры

Нам дана обучающая выборка

$$X_{tr}, T_{tr} = \{x_i, t_i\}_{i=1}^n$$

Хотим оценить *параметры модели* w .



Нахождение параметров

Q: Как находить параметры w ?

Нахождение параметров

Q: Как находить параметры w ?

A: Априорного распределения нет — это ещё не байесовские методы.

Значит, нужно использовать **метод максимума правдоподобия**.

Нахождение параметров

Q: Как находить параметры w ?

A: Априорного распределения нет — это ещё не байесовские методы.

Значит, нужно использовать **метод максимума правдоподобия**.

$$w_{ML} = \underset{w}{\operatorname{argmax}} p(T_{tr} | X_{tr}, w) = \underset{w}{\operatorname{argmax}} \prod_{i=1}^n p(t_i | x_i, w).$$

Естественно, от произведения сразу переходим к сумме логарифмов:

$$w_{ML} = \underset{w}{\operatorname{argmax}} \sum_{i=1}^n \log p(t_i | x_i, w).$$

Метод максимум правдоподобия

Напомним, что

$$p(t | x, w) = \mathcal{N}(t | w^T x, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2} (t - x^T w)^2\right).$$

Подставляем в

$$w_{ML} = \operatorname{argmax}_w \sum_{i=1}^n \log p(t_i | x_i, w).$$

Метод максимум правдоподобия

Напомним, что

$$p(t | x, w) = \mathcal{N}(t | w^T x, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2} (t - x^T w)^2\right).$$

Подставляем в

$$w_{ML} = \underset{w}{\operatorname{argmax}} \sum_{i=1}^n \log p(t_i | x_i, w).$$

Убираем константы (β не влияет на argmax):

$$w_{ML} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^n (t_i - x_i^T w)^2 = \underset{w}{\operatorname{argmin}} \|T_{tr} - X_{tr} w\|^2.$$

Q: Какое здесь решение?

Получаем известный ответ:

Теорема

MLE — это псевдорешение:

$$\mathbf{w}_{ML} = \left(\mathbf{X}_{tr}^T \mathbf{X}_{tr} \right)^{-1} \mathbf{X}_{tr}^T \mathbf{T}_{tr}.$$

Чтобы было удобнее следить за размерами матриц: \mathbf{T}_{tr} — это $n \times 1$ матрица, \mathbf{X}_{tr} — это $n \times m$ матрица и \mathbf{w} — это $m \times 1$ матрица.

Линейная регрессия по-байесовски

1 Линейная регрессия по-статистически

2 Линейная регрессия по-байесовски

- МАР-оценка
- Аналитический байесовский вывод

3 Метод релевантных векторов

- Вариационная нижняя оценка
- Алгоритм

Байес возвращается

Переходим к байесовским методам.

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

posterior prior likelihood marginal



Вводим априорное распределение на параметры $p(w)$.

Всё будет нормально!

Естественно, всё будет нормально $p(w) \sim \mathcal{N}(w | 0, \alpha^{-1}I)$.

Матрица ковариации сейчас — скалярная

$$\Sigma = \begin{pmatrix} \frac{1}{\alpha} & & \\ & \ddots & \\ & & \frac{1}{\alpha} \end{pmatrix}.$$

Q: Как теперь выглядит наша модель?

Было распределение $p(t | x, w)$, а теперь — какое?

Что делать?

Мы рассматриваем распределение

$$\begin{aligned} p(t, w | x) &= p(t | x, w)p(w) \\ &= \mathcal{N}(t | w^T x, \beta^{-1})\mathcal{N}(w | 0, \alpha^{-1}I). \end{aligned}$$

x — фиксировано. Мы его не предсказываем и не моделируем.

Q: Что мы можем сделать?

Ближайшие действия

Дана обучающая выборка:

$$X_{tr}, T_{tr} = \{x_i, t_i\}_{i=1}^n$$

План работ:

- ① Найдём МАР-оценку

$$w_{MP} = \underset{w}{\operatorname{argmax}} p(w | X_{tr}, T_{tr}).$$

- ② Вычислим явно апостериорное распределение

$$p(w | X_{tr}, T_{tr}).$$

MAP-оценка

1 Линейная регрессия по-статистически

2 Линейная регрессия по-байесовски

- MAP-оценка
- Аналитический байесовский вывод

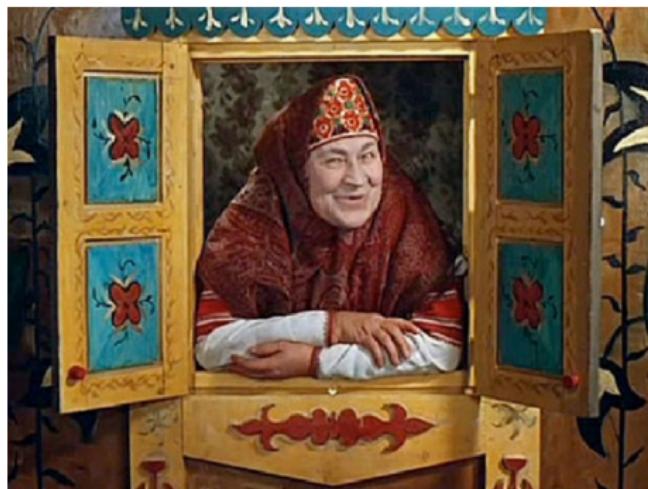
3 Метод релевантных векторов

- Вариационная нижняя оценка
- Алгоритм

Простой план

Ищем МАР-оценку $\underset{w}{\operatorname{argmax}} p(w | X_{tr}, T_{tr})$. Идеино всё просто.

- ➊ Применяя формулу Байеса.
- ➋ Подставляем известные распределения $p(t | x, w)p(w)$ и честно считаем.



Скоро сказка сказывается
да нескоро дело делается

Формула Байеса

- Ищем моду апостериорного. По формуле Байеса:

$$w_{MP} = \operatorname{argmax}_w p(w | X_{tr}, T_{tr}) = \operatorname{argmax}_w p(T_{tr} | X_{tr}, w)p(w)$$

Q: Куда пропал знаменатель в теореме Байеса?

Формула Байеса

- Ищем моду апостериорного. По формуле Байеса:

$$w_{MP} = \operatorname{argmax}_w p(w | X_{tr}, T_{tr}) = \operatorname{argmax}_w p(T_{tr} | X_{tr}, w)p(w)$$

Q: Куда пропал знаменатель в теореме Байеса?

A: Он не влияет на argmax.

- Далее, как обычно, логарифмируем произведение.

Правдоподобие $p(T_{tr} | X_{tr}, w)$ распадётся в произведение правдоподобий по отдельным объектам:

$$w_{MP} = \operatorname{argmax}_w \left(\sum_{i=1}^n [\ln p(t_i | x_i, w)] + \ln p(w) \right)$$

Нормальные распределения

- Подставляем формулы для нормального распределения:

$$\begin{aligned} p(t | x, w)p(w) &= \mathcal{N}(t | w^T x, \beta^{-1}) \mathcal{N}(w | 0, \alpha^{-1} I) = \\ &= \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2} (t - x^T w)^2\right) \left(\frac{\alpha}{2\pi}\right)^{\frac{d}{2}} \exp\left(-\frac{\alpha}{2} w^T w\right). \end{aligned}$$

Нормальные распределения

- Подставляем формулы для нормального распределения:

$$\begin{aligned} p(t | x, w)p(w) &= \mathcal{N}(t | w^T x, \beta^{-1}) \mathcal{N}(w | 0, \alpha^{-1} I) = \\ &= \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2} (t - x^T w)^2\right) \left(\frac{\alpha}{2\pi}\right)^{\frac{d}{2}} \exp\left(-\frac{\alpha}{2} w^T w\right). \end{aligned}$$

- Логарифмируем и подставляем в argmax:

$$w_{MP} = \underset{w}{\operatorname{argmax}} \left[\frac{n}{2} \log \beta - \frac{\beta}{2} \sum_{i=1}^n (t_i - x_i^T w)^2 + \frac{d}{2} \log \alpha - \frac{\alpha}{2} w^T w + const \right].$$

Синее не зависит от w .

Argmax

- Отбрасываем то, что не зависит от w , и записываем всё в матричном виде:

$$w_{MP} = \underset{w}{\operatorname{argmax}} \left[-\frac{\beta}{2} (T_{tr} - X_{tr} w)^T (T_{tr} - X_{tr} w) - \frac{\alpha}{2} w^T w \right].$$

Argmax

- Отбрасываем то, что не зависит от w , и записываем всё в матричном виде:

$$w_{MP} = \underset{w}{\operatorname{argmax}} \left[-\frac{\beta}{2} (T_{tr} - X_{tr}w)^T (T_{tr} - X_{tr}w) - \frac{\alpha}{2} w^T w \right].$$

- Теперь находим экстремум. Нужно приравнивать производную по w к нулю.

Если всё правильно посчитать:

$$\beta X_{tr}^T T_{tr} - (\beta X_{tr}^T X_{tr} + \alpha I) w = 0,$$

MAP-оценка

$$w_{MP} = \beta \left(\beta X_{tr}^T X_{tr} + \alpha I \right)^{-1} X_{tr}^T T_{tr}.$$

MAP-оценка

MAP-оценка

$$w_{MP} = \beta \left(\beta X_{tr}^T X_{tr} + \alpha I \right)^{-1} X_{tr}^T T_{tr}.$$

Похоже на то, что было раньше:

$$w_{ML} = \left(X_{tr}^T X_{tr} \right)^{-1} X_{tr}^T T_{tr}.$$

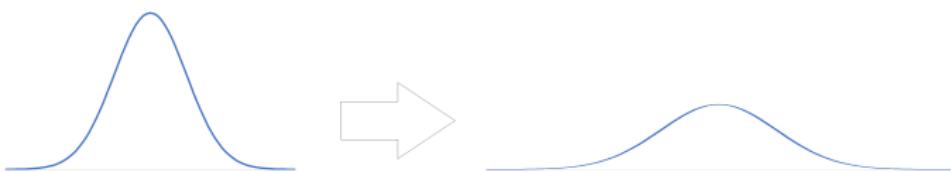
Только слагаемое αI в скобках появилось.

Параметр модели

- Пусть $\alpha \rightarrow 0$.

$$\lim_{\alpha \rightarrow 0} \omega_{MP} = \omega_{ML}.$$

Априорное распределение становится всё шире и шире, почти константой.



$p(w)$ предсказумо перестаёт влиять на argmax.

Параметр модели

- Пусть $\alpha \rightarrow \infty$. Тогда

$$\omega_{MP} \approx \frac{\beta}{\alpha} X_{tr}^T T_{tr}.$$

Априорное распределение схлопывается в дельта-функцию.

Неопределённость на w исчезает:

$$\lim_{\alpha \rightarrow +\infty} \omega_{MP} = 0.$$

Перерыв

Перерыв

MAP-оценка

1 Линейная регрессия по-статистически

2 Линейная регрессия по-байесовски

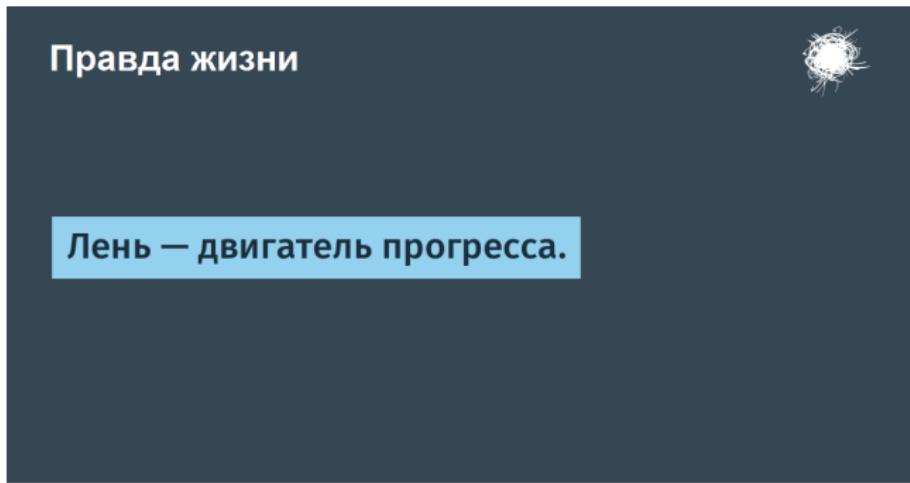
- MAP-оценка
- Аналитический байесовский вывод

3 Метод релевантных векторов

- Вариационная нижняя оценка
- Алгоритм

Аналитический байесовский вывод

Будем честно вычислять апостериорное распределение $p(w | X_{tr}, T_{tr})$?



Всё будет нормально!

Q: А мы вообще можем выполнить аналитический байесовский вывод?

$$p(t | x, w)p(w) = \mathcal{N}(t | w^T x, \beta^{-1})\mathcal{N}(w | 0, \alpha^{-1}I)$$

Всё будет нормально!

Q: А мы вообще можем выполнить аналитический байесовский вывод?

$$p(t | x, w)p(w) = \mathcal{N}(t | w^T x, \beta^{-1})\mathcal{N}(w | 0, \alpha^{-1}I)$$

A: Да, и априорное распределение, и функция правдоподобия — “перевёрнутые параболы под экспонентой”. Функциональный вид сохраняется:

$$\exp\left(-\frac{w^T \Sigma_1 w}{2} + \dots\right) \exp\left(-\frac{w^T \Sigma_2 w}{2} + \dots\right) = \exp\left(-\frac{w^T (\Sigma_1 + \Sigma_2) w}{2} + \dots\right)$$

Апостериорное распределение

Q: Итак, какой вид имеет $p(w | X_{tr}, T_{tr})$?

Апостериорное распределение

Q: Итак, какой вид имеет $p(w | X_{tr}, T_{tr})$?

A: Это нормальное распределение

$$p(w | X_{tr}, T_{tr}) \sim \mathcal{N}(w | \mu, \Sigma)$$

Параметры нормального распределения

Q: Как найти параметры μ и Σ нормального распределения?

$$\mathcal{N}(w | \mu, \Sigma) = \left(\frac{1}{2\pi} \right)^{\frac{d}{2}} \exp \left(-\frac{1}{2} (w - \mu)^T \Sigma^{-1} (w - \mu) \right)$$

Параметры нормального распределения

Q: Как найти параметры μ и Σ нормального распределения?

$$\mathcal{N}(w | \mu, \Sigma) = \left(\frac{1}{2\pi} \right)^{\frac{d}{2}} \exp \left(-\frac{1}{2} (w - \mu)^T \Sigma^{-1} (w - \mu) \right)$$

A:

- μ — это мода (точка максимума)
- Σ — коэффициент в квадратичной части под экспонентой $-\frac{1}{2} w^T \Sigma^{-1} w$.

Мода

Q: Как найти моду μ ?

Мода

Q: Как найти моду μ ?

A: А мы её уже нашли — это МАР-оценка:

$$\mu = w_{MP} = \beta \left(\beta X_{tr}^T X_{tr} + \alpha I \right)^{-1} X_{tr}^T T_{tr}.$$

Матрица ковариации

Q: Как найти Σ ? Мы уже выделили слагаемые с w при поиске argmax:

$$w_{MP} = \underset{w}{\operatorname{argmax}} \left[-\frac{\beta}{2} (T_{tr} - X_{tr}w)^T (T_{tr} - X_{tr}w) - \frac{\alpha}{2} w^T w \right].$$

Остается выделить слагаемое вида $-\frac{1}{2} w^T \Sigma^{-1} w$.

Матрица ковариации

Q: Как найти Σ ? Мы уже выделили слагаемые с w при поиске argmax:

$$w_{MP} = \underset{w}{\operatorname{argmax}} \left[-\frac{\beta}{2} (T_{tr} - X_{tr}w)^T (T_{tr} - X_{tr}w) - \frac{\alpha}{2} w^T w \right].$$

Остается выделить слагаемое вида $-\frac{1}{2} w^T \Sigma^{-1} w$.

Раскрываем скобки — видим, что

$$\Sigma = (\beta X_{tr}^T X_{tr} + \alpha I)^{-1}$$

Теорема

Рассмотрим линейную модель с гауссовским шумом:

$$p(t | x, w) \sim \mathcal{N}(t | w^T x, \beta^{-1}).$$

Если априорное распределение имеет нормальное распределение вида

$$p(w) \sim \mathcal{N}(w | 0, \alpha^{-1} I),$$

то апостериорное распределение тоже нормально

$$p(w | X_{tr}, T_{tr}) \sim \mathcal{N}(w | \mu, \Sigma).$$

Априорное распределение. Итоги

Теорема

Рассмотрим линейную модель с гауссовским шумом:

$$p(t | x, w) \sim \mathcal{N}(t | w^T x, \beta^{-1}).$$

Если априорное распределение имеет нормальное распределение вида

$$p(w) \sim \mathcal{N}(w | 0, \alpha^{-1} I),$$

то апостериорное распределение тоже нормально

$$p(w | X_{tr}, T_{tr}) \sim \mathcal{N}(w | \mu, \Sigma).$$

- Мода:

$$\mu = w_{MP} = \beta \left(\beta X_{tr}^T X_{tr} + \alpha I \right)^{-1} X_{tr}^T T_{tr}.$$

- Матрица ковариации:

$$\Sigma = \left(\beta X_{tr}^T X_{tr} + \alpha I \right)^{-1}$$

Подбор гиперпараметров

У нас есть 2 параметра α и β .

Q: Как подобрать их по данным?

Подбор гиперпараметров

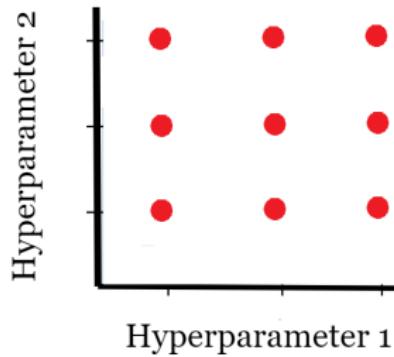
У нас есть 2 параметра α и β .

Q: Как подобрать их по данным?

A: Проще всего — [кросс-валидацией](#). Пока параметров 2, это ещё выполнимо.

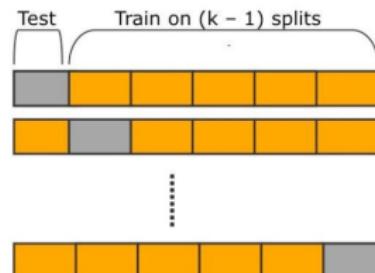
GridSearch

"Перебираем параметры по сеточке"



Cross-Validation

"Обучаемся на $n-1$ элементе, проверяем на оставшемся"



Перерыв

Перерыв

Метод релевантных векторов

1 Линейная регрессия по-статистически

2 Линейная регрессия по-байесовски

- МАР-оценка
- Аналитический байесовский вывод

3 Метод релевантных векторов

- Вариационная нижняя оценка
- Алгоритм

Feature selection

Вспомним задачу отбора признаков (Feature selection).

Есть основания полагать, что многие признаки — шумовые.

All Features



Feature Selection



Final Features



One size doesn't fit all

Параметр α отвечает за “надстройку под параметры”:

- $\alpha \rightarrow 0 \Rightarrow$ больше подстраиваемся под признак.
- $\alpha \rightarrow \infty \Rightarrow$ меньше учитываем признак.

Q: Наверное, один параметр α на все признаки — не лучшая идея, да?

ONE SIZE
DOESN'T FIT ALL



Куча параметров

Идея! Каждому признаку — свой параметр α_i !

Заменяем матрицу ковариации $\alpha^{-1}I$ на диагональную матрицу

$$A^{-1} = \begin{pmatrix} \frac{1}{\alpha_1} & & \\ & \ddots & \\ & & \frac{1}{\alpha_d} \end{pmatrix}.$$

Новое распределение

Итак, новое распределение:

$$p(t, w | x) = p(t | x, w, \beta)p(w | A) = \mathcal{N}(t | w^T x, \beta^{-1})\mathcal{N}(w | 0, A^{-1}).$$

Q: Итак, $\alpha I \rightarrow A$. Как думаете — что произойдёт с апостериорным распределением?

Новое апостериорное

- Мода:

$$\mu = w_{MP} = \beta \left(\beta X_{tr}^T X_{tr} + A \right)^{-1} X_{tr}^T T_{tr}.$$

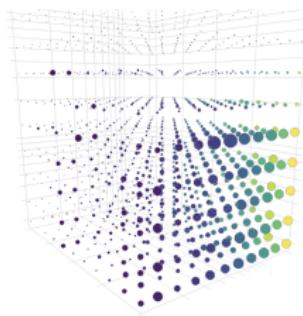
- Матрица ковариации:

$$\Sigma = \left(\beta X_{tr}^T X_{tr} + A \right)^{-1}$$

Проклятие размерности

Q: Как найти параметры A и β модели?

Кросс-валидация не годится! Слишком много параметров — $O(2^d)$ вариантов!



Принцип наибольшей обоснованности

- Воспользуемся принципом наибольшей обоснованности для выбора параметров:

$$(A, \beta) = \underset{A, \beta}{\operatorname{argmax}} p(T_{tr} | X_{tr}, A, \beta).$$

Принцип наибольшей обоснованности

- Воспользуемся принципом наибольшей обоснованности для выбора параметров:

$$(A, \beta) = \operatorname{argmax}_{A, \beta} p(T_{tr} | X_{tr}, A, \beta).$$

- Полностью выписываем модель, маргинализируя по всем неявным параметрам:

$$(A, \beta) = \operatorname{argmax}_{A, \beta} \int p(T_{tr}, w | X_{tr}, A, \beta) dw$$

Принцип наибольшей обоснованности

- Воспользуемся принципом наибольшей обоснованности для выбора параметров:

$$(A, \beta) = \operatorname{argmax}_{A, \beta} p(T_{tr} | X_{tr}, A, \beta).$$

- Полностью выписываем модель, маргинализируя по всем неявным параметрам:

$$(A, \beta) = \operatorname{argmax}_{A, \beta} \int p(T_{tr}, w | X_{tr}, A, \beta) dw$$

- Применяем правило произведения

$$(A, \beta) = \operatorname{argmax}_{A, \beta} \int p(T_{tr} | X_{tr}, w, \beta) p(w | A) dw.$$

Спойлеры

Здесь можно сразу перейти в **конец презентации** — там будет ответ.

Спойлер. Итоговое решение — итерационный процесс вида

$$\alpha^{j+1} = f(\alpha^j, \beta^j, w^j), \quad \beta^{j+1} = g(\alpha^j, \beta^j, w^j), \quad w^{j+1} = h(\alpha^j, \beta^j, w^j)$$

Оставшаяся часть лекции будет объяснением того, “как мы дошли до жизни такой”.



Рис.: Настало время срезать углы

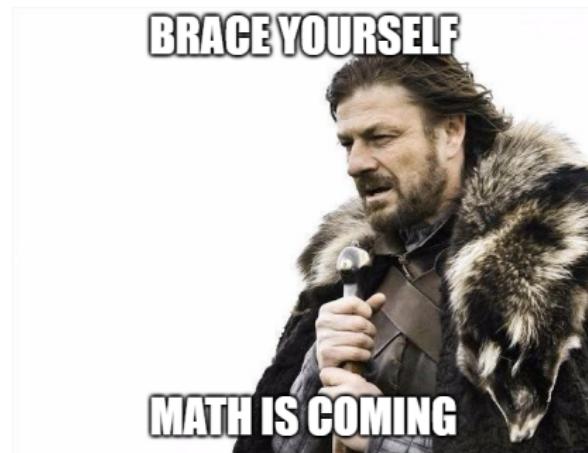
Brace yourself

Нам нужно посчитать

$$(A, \beta) = \underset{A, \beta}{\operatorname{argmax}} \int p(T_{tr} | X_{tr}, w, \beta) p(w | A) dw.$$

У нас всё нормально \Rightarrow под интегралом — экспонента от квадратичной функции.

Всё считается, и так формулы сложные — сразу напишем ответ.



Функция под логарифмом

Подынтегральное выражение

$$p(T_{tr} | X_{tr}, w, \beta) p(w | A) = \mathcal{N}(T_{tr} | X_{tr}w, \beta^{-1}) \mathcal{N}(w | 0, A^{-1})$$

Функция под логарифмом

Подынтегральное выражение

$$p(T_{tr} | X_{tr}, w, \beta) p(w | A) = \mathcal{N}(T_{tr} | X_{tr}w, \beta^{-1}) \mathcal{N}(w | 0, A^{-1})$$

Честно его считаем и подставляем:

$$\left(\frac{\beta}{2\pi}\right)^{\frac{n}{2}} \exp\left[-\frac{\beta}{2} (T_{tr} - X_{tr}w)^2\right] \left(\frac{1}{2\pi}\right)^{\frac{d}{2}} \sqrt{\det A} \exp\left[-\frac{1}{2} w^T A w\right].$$

Функция под логарифмом

Подынтегральное выражение

$$p(T_{tr} | X_{tr}, w, \beta) p(w | A) = \mathcal{N}(T_{tr} | X_{tr}w, \beta^{-1}) \mathcal{N}(w | 0, A^{-1})$$

Честно его считаем и подставляем:

$$\left(\frac{\beta}{2\pi}\right)^{\frac{n}{2}} \exp\left[-\frac{\beta}{2} (T_{tr} - X_{tr}w)^2\right] \left(\frac{1}{2\pi}\right)^{\frac{d}{2}} \sqrt{\det A} \exp\left[-\frac{1}{2} w^T A w\right].$$

Можно проверить, что

$$\begin{aligned} \log p(T_{tr} | X_{tr}, A, \beta) &= \frac{n}{2} \log \beta - \frac{\beta}{2} \|T_{tr} - X_{tr}w_{MP}\|^2 + \\ &+ \frac{1}{2} \log \det A - \frac{1}{2} w_{MP}^T A w_{MP} + \\ &+ \frac{1}{2} \log \det \Sigma + \text{const.} \end{aligned}$$

Неудобно дифференцировать

Хотим дифференцировать

$$\begin{aligned}\log p(T_{tr} | X_{tr}, A, \beta) = & \frac{n}{2} \log \beta - \frac{\beta}{2} \|T_{tr} - X_{tr} w_{MP}\|^2 + \\ & + \frac{1}{2} \log \det A - \frac{1}{2} w_{MP}^T A w_{MP} + \\ & + \frac{1}{2} \log \det \Sigma + \text{const.}\end{aligned}$$

по A и β и приравнять производные к нулю.

Неудобно дифференцировать

Хотим дифференцировать

$$\begin{aligned}\log p(T_{tr} | X_{tr}, A, \beta) = & \frac{n}{2} \log \beta - \frac{\beta}{2} \|T_{tr} - X_{tr} w_{MP}\|^2 + \\ & + \frac{1}{2} \log \det A - \frac{1}{2} w_{MP}^T A w_{MP} + \\ & + \frac{1}{2} \log \det \Sigma + \text{const.}\end{aligned}$$

по A и β и приравнять производные к нулю.

Проблема. Последнее слагаемое дифференцировать неудобно:

$$\Sigma = (\beta X_{tr}^T X_{tr} + A)^{-1}$$

Q: Что будем делать?

Неудобно дифференцировать

Хотим дифференцировать

$$\begin{aligned}\log p(T_{tr} | X_{tr}, A, \beta) &= \frac{n}{2} \log \beta - \frac{\beta}{2} \|T_{tr} - X_{tr} w_{MP}\|^2 + \\ &+ \frac{1}{2} \log \det A - \frac{1}{2} w_{MP}^T A w_{MP} + \\ &+ \frac{1}{2} \log \det \Sigma + \text{const.}\end{aligned}$$

по A и β и приравнять производные к нулю.

Проблема. Последнее слагаемое дифференцировать неудобно:

$$\Sigma = (\beta X_{tr}^T X_{tr} + A)^{-1}$$

Q: Что будем делать?

A: Меняем

$$\frac{1}{2} \log \det \Sigma = -\frac{1}{2} \log \det \Sigma^{-1}$$

Не считается

Выделили слагаемые с A и β :

$$\begin{aligned}\log p(T_{tr} | X_{tr}, A, \beta) &= \frac{n}{2} \log \beta - \frac{\beta}{2} \|T_{tr} - X_{tr}^T w_{MP}\|^2 + \\ &+ \frac{1}{2} \log \det A - \frac{1}{2} w_{MP}^T A w_{MP} + \\ &+ \frac{1}{2} \log \det \Sigma + \text{const.}\end{aligned}$$

Q: Ничего не забыли?

Не считается

Выделили слагаемые с A и β :

$$\begin{aligned}\log p(T_{tr} | X_{tr}, A, \beta) &= \frac{n}{2} \log \beta - \frac{\beta}{2} \|T_{tr} - X_{tr}^T w_{MP}\|^2 + \\ &+ \frac{1}{2} \log \det A - \frac{1}{2} w_{MP}^T A w_{MP} + \\ &+ \frac{1}{2} \log \det \Sigma + \text{const.}\end{aligned}$$

Q: Ничего не забыли?

A: Теперь w_{MP} зависит от A :

$$w_{MP} = \beta \left(\beta X_{tr}^T X_{tr} + A \right)^{-1} X_{tr}^T T_{tr}$$

Проблема: Если бы вместо w_{MP} было w , то могли бы оптимизировать. А так — сложно!

Если очень хочется...

Мы оптимизируем функцию вида

$$F(A, \beta, w_{MP})$$

Идея. Заменим w_{MP} на w , не зависящую от A, b . Для функции

$$F(A, \beta, w)$$

мы сможем аналитически решить задачу.

«Если нельзя,
но очень хочется,
то можно...»

Новая оптимизационная задача

Остается понять — насколько мы ошибаемся.

По определению w_{MP} — точка, где достигается максимум

$$w_{MP} = \underset{w}{\operatorname{argmax}} F(A, \beta, w).$$

Далее посмотрим — какая оптимизационная задача у нас возникает.

Перерыв

Перерыв

Метод релевантных векторов

1 Линейная регрессия по-статистически

2 Линейная регрессия по-байесовски

- МАР-оценка
- Аналитический байесовский вывод

3 Метод релевантных векторов

- Вариационная нижняя оценка
- Алгоритм

ELBO

Функция $g(x, \xi)$ — вариационная нижняя оценка (ELBO) на функцию $f(x)$, если

- ① для всех x, ξ выполнено $f(x) \geq g(x, \xi)$.
- ② для любого x_0 существует ξ_0 т.ч. $f(x_0) = g(x_0, \xi_0)$.

ELBO

Функция $g(x, \xi)$ — вариационная нижняя оценка (ELBO) на функцию $f(x)$, если

- ① для всех x, ξ выполнено $f(x) \geq g(x, \xi)$.
- ② для любого x_0 существует ξ_0 т.ч. $f(x_0) = g(x_0, \xi_0)$.

Проще говоря:

- $f(x)$ всегда “выше” $g(x, \xi)$.
- на каждой вертикали $x = x_0$ семейство $g(x, \xi)$ “доходит” до $f(x)$.

ELBO

По сути $f(x)$ — огибающая семейства $g(x, \xi)$.

Грубо говоря, если нарисовать все $g(x, \xi)$, то мы “увидим сверху” $f(x)$.

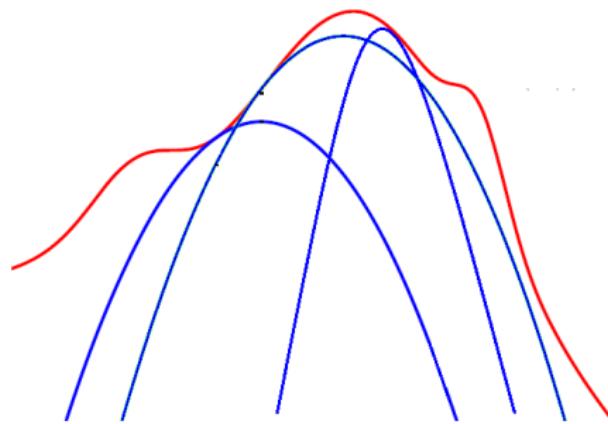
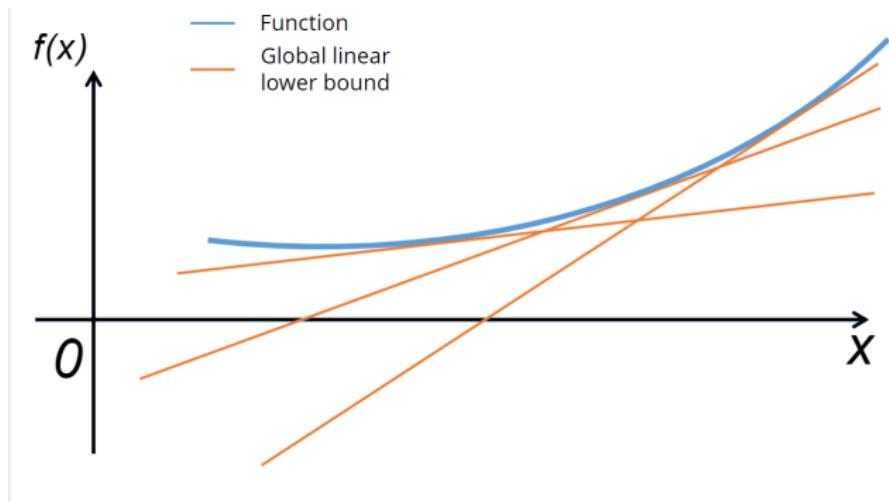


Рис.: Синее семейство — вариационная нижняя оценка функции $f(x)$

ELBO

Q: Знакомы ли мы с какими-нибудь примерами вариационной нижней оценки (ELBO)?

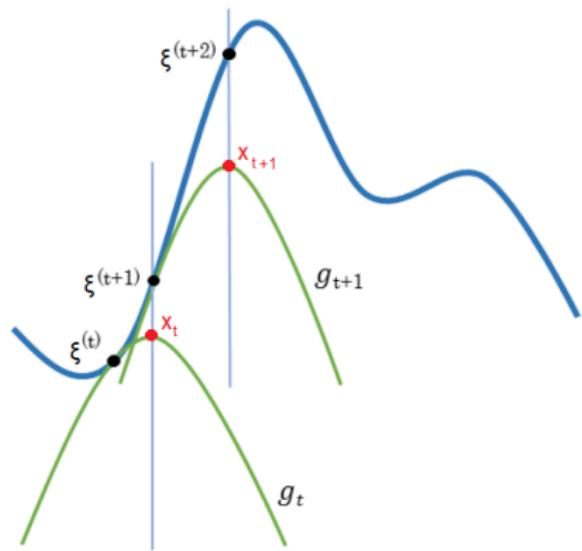
A: Пример ELBO: выпуклая функция — огибающая своего семейства касательных.



EM-алгоритм

Q: Как максимизировать $f(x)$? A: Можно по очереди максимизировать:

- Брать максимум “по горизонтали” (по x) у очередной функции g_t .
- Брать максимум “по вертикали” (по ξ).



EM-алгоритм

EM-алгоритм

Итеративный процесс для поиска (локального) максимума:

$$\begin{cases} x_{n+1} = \underset{x}{\operatorname{argmax}} g(x, \xi_n) \\ \xi_{n+1} = \underset{\xi}{\operatorname{argmax}} g(x_{n+1}, \xi) \end{cases}$$

Алгоритм

1 Линейная регрессия по-статистически

2 Линейная регрессия по-байесовски

- МАР-оценка
- Аналитический байесовский вывод

3 Метод релевантных векторов

- Вариационная нижняя оценка
- Алгоритм

Оптимизация по параметрам

Итак, заменяем w_{MP} на w . Максимизируем по (A, β) и ещё по w выражение:

$$\begin{aligned}\log p(T_{tr} | X_{tr}, A, \beta) = & \frac{n}{2} \log \beta - \frac{\beta}{2} \|T_{tr} - X_{tr}w\|^2 + \\ & + \frac{1}{2} \log \det A - \frac{1}{2} w^T A w + \\ & - \frac{1}{2} \log \det \Sigma^{-1} + \text{const.}\end{aligned}$$

Оптимизация по параметрам

Итак, заменяем w_{MP} на w . Максимизируем по (A, β) и ещё по w выражение:

$$\begin{aligned}\log p(T_{tr} | X_{tr}, A, \beta) = & \frac{n}{2} \log \beta - \frac{\beta}{2} \|T_{tr} - X_{tr}w\|^2 + \\ & + \frac{1}{2} \log \det A - \frac{1}{2} w^T A w + \\ & - \frac{1}{2} \log \det \Sigma^{-1} + \text{const.}\end{aligned}$$

Для примера продифференцируем по α_j — элементу A (он есть в красных слагаемых). Можно проверить, что получится

$$\frac{1}{2\alpha_j} - \frac{1}{2} w_j^2 - \frac{1}{2} \Sigma_{jj} = 0.$$

Опыт показал

Чтобы процесс лучше сходился, лучше оптимизировать $\log \alpha_j$, а не α_j .

Домножим выражение

$$\frac{1}{2\alpha_j} - \frac{1}{2} w_j^2 - \frac{1}{2} \Sigma_{jj} = 0.$$

на $2\alpha_j$ и представим в виде:

$$\alpha_j w_j^2 = 1 - \alpha_j \Sigma_{jj}$$

Метод простой итерации

Уравнения $x - f(x) = 0$ можно решать методом простой итерации

$$x_{n+1} = f(x_n).$$

Отметим, что Σ_{jj} также зависит от α_j . Получаем итеративный процесс

$$\alpha_j^{new} = \frac{1 - \alpha_j^{old} \Sigma_{jj}^{old}}{w_j^2}.$$

Для β и w формулы получаются аналогично.

Relevance Vector Machine

Итеративный процесс обновления параметров α, β, w :

- Обновляем α :

$$\alpha_j^{new} = \frac{1 - \alpha_j^{old} \Sigma_{jj}^{old}}{w_j^2}.$$

- Затем β :

$$\beta_j^{new} = \frac{n - \sum_{j=1}^d (1 - \alpha_j^{old} \Sigma_{jj}^{old})}{\|T_{tr} - X_{tr} w_j\|^2}.$$

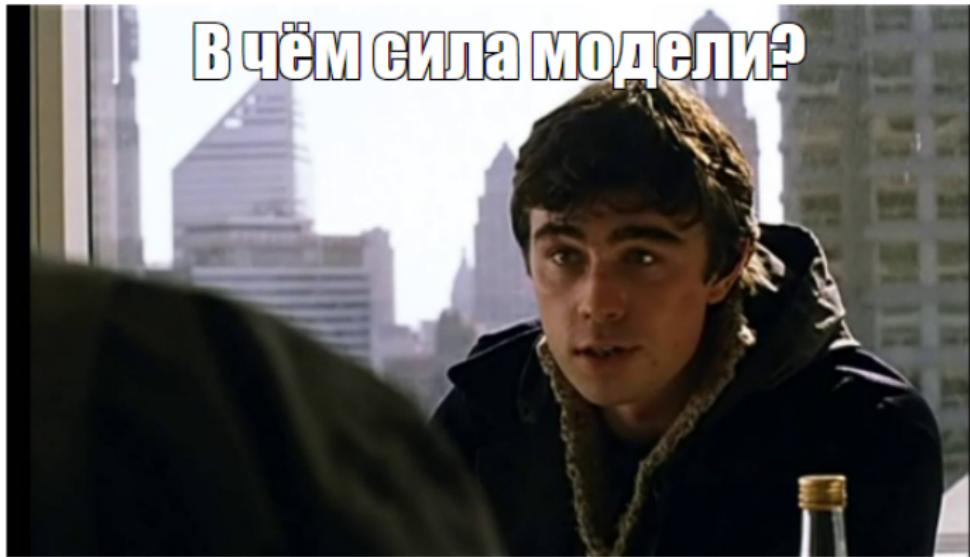
- И наконец w :

$$w_j^{new} = \beta^{new} \left(\beta^{new} X_{tr}^T X_{tr} + A_{new} \right)^{-1} X_{tr}^T T_{tr}$$

Повторить процесс несколько раз (до сходимости).

Преимущество модели

И чем эта модель так хороша?



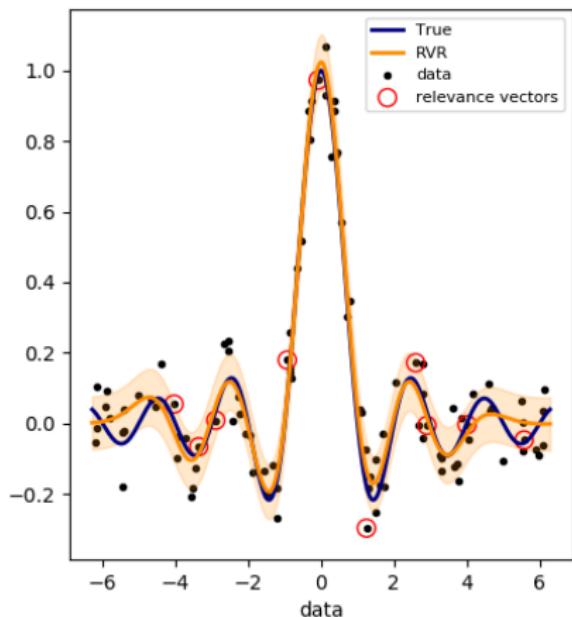
В чём сила модели?

Преимущество модели

Ещё один способ отбора признаков.

Если $\alpha_i \rightarrow \infty$ — признак шумовой.

Остаются только самые важные признаки.



We did it!



мы строили, строили
и наконец построили! ура-а-а!!