

Метод релевантных векторов

Д.П. Ветров¹ Д.А. Кропотов²

¹МГУ, ВМиК, каф. ММП

²ВЦ РАН

Спецкурс «Байесовские методы машинного обучения»

План лекции

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- 1 Ликбез
- 2 Метод релевантных векторов для задачи регрессии
- 3 Метод релевантных векторов для задачи классификации

Матричные тождества

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- Тожество Шермана-Моррисона-Вудбери

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

Здесь $A \in \mathbb{R}_{n \times n}$, $U \in \mathbb{R}_{n \times m}$, $C \in \mathbb{R}_{m \times m}$, $V \in \mathbb{R}_{m \times n}$. Если нам известно A^{-1} и $m \ll n$, то обращать матрицу $C^{-1} + VA^{-1}U$ проще, чем $A + UCV$.

- Лемма об определителе матрицы

$$\det(A + UV) = \det(I + VA^{-1}U) \det(A)$$

Здесь $A \in \mathbb{R}_{n \times n}$, $U \in \mathbb{R}_{n \times m}$, $V \in \mathbb{R}_{m \times n}$, I — единичная матрица размера $m \times m$.

Тождество Шермана-Моррисона-Вудбери

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- Тождество

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

- Доказательство

$$\begin{aligned}(A + UCV)(A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}) &= \\ I + UCVA^{-1} - (U + UCVA^{-1}U)(C^{-1} + VA^{-1}U)^{-1}VA^{-1} &= \\ I + UCVA^{-1} - UC\cancel{(C^{-1} + VA^{-1}U)}\cancel{(C^{-1} + VA^{-1}U)}^{-1}VA^{-1} &= \\ I + UCVA^{-1} - UCVA^{-1} &= I\end{aligned}$$

Тождества для определителей матрицы

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- При доказательстве многих матричных тождеств полезным оказывается следующее равенство:

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(A) \det(D - CA^{-1}B) = \det(D) \det(A - BD^{-1}C)$$

Здесь $A \in \mathbb{R}_{m \times m}$, $B \in \mathbb{R}_{m \times n}$, $C \in \mathbb{R}_{n \times m}$, $D \in \mathbb{R}_{n \times n}$

- Это равенство следует из следующего тождества:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A & 0 \\ C & I \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & D - CA^{-1}B \end{pmatrix} = \begin{pmatrix} I & B \\ 0 & D \end{pmatrix} \begin{pmatrix} A - BD^{-1}C & 0 \\ D^{-1}C & I \end{pmatrix}$$

- Лемма об определителе матрицы

$$\det(A + UV) = \det(I + VA^{-1}U) \det(A)$$

- Доказательство:

$$\det \begin{pmatrix} A & -U \\ V & I \end{pmatrix} = \det(A) \det(I + VA^{-1}U) =$$

$$\det(I) \det(A + UI^{-1}V) = \det(A + UV)$$

Обобщенные линейные модели (напоминание)

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- Рассмотрим следующую задачу восстановления регрессии: имеется выборка $(X, t) = \{\mathbf{x}_i, t_i\}_{i=1}^n$, где вектор признаков $\mathbf{x}_i \in \mathbb{R}^d$, а целевая переменная $t_i \in \mathbb{R}$, требуется для нового объекта \mathbf{x}_* предсказать значение целевой переменной t_* .
- Предположим, что $t = f(\mathbf{x}) + \varepsilon$, где $\varepsilon \sim \mathcal{N}(\varepsilon|0, \sigma^2)$, а

$$f(\mathbf{x}) = \sum_{j=1}^m w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

Здесь \mathbf{w} — набор числовых параметров, а $\boldsymbol{\phi}(\mathbf{x})$ — вектор обобщенных признаков.

- Часто в качестве обобщенных признаков выбираются следующие:
 - Обычные признаки — $\phi_j(\mathbf{x}) = x_j, j = 1, \dots, d$
 - Ядровые функции —
 $\phi_j(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}_j), j = 1, \dots, n, \phi_{n+1}(\mathbf{x}) \equiv 1$

Метод максимума правдоподобия (линейная регрессия)

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- Так как шумовая компонента ε имеет независимое нормальное распределение, то можно выписать функцию правдоподобия обучающей выборки:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^n p(t_i|\mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^n \mathcal{N}(t_i|f(\mathbf{x}_i, \mathbf{w}), \beta^{-1}) = \prod_{i=1}^n \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}(t_i - \mathbf{w}^T \phi(\mathbf{x}_i))^2\right) = \sqrt{\frac{\beta}{2\pi}}^n \exp\left(-\frac{\beta}{2} \sum_{i=1}^n (t_i - \mathbf{w}^T \phi(\mathbf{x}_i))^2\right)$$

- Переходя к логарифму, получаем

$$-\frac{\beta}{2} \sum_{i=1}^n (t_i - \mathbf{w}^T \phi(\mathbf{x}_i))^2 = -\frac{\beta}{2} (\mathbf{t} - \Phi \mathbf{w})^T (\mathbf{t} - \Phi \mathbf{w}) \rightarrow \max_{\mathbf{w}}$$

Здесь $\Phi = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_n)]^T$.

- Точка максимума правдоподобия выписывается в явном виде:

$$\mathbf{w}_{ML} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t}$$

Введение регуляризации (априорного распределения)

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- Следуя байесовскому подходу воспользуемся методом максимума апостериорной плотности:

$$w_{MP} = \arg \max_w p(\mathbf{w}|X, \mathbf{t}) = \arg \max_w p(\mathbf{t}|X, \mathbf{w})p(\mathbf{w})$$

- Выберем в качестве априорного распределения на параметры \mathbf{w} следующее:

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

Такой выбор соответствует штрафу за большие значения коэффициентов \mathbf{w} с параметром регуляризации α .

- Максимизация апостериорной плотности эквивалентна следующей задаче оптимизации:

$$-\frac{\beta}{2} \sum_{i=1}^n (t_i - \mathbf{w}^T \phi(\mathbf{x}_i))^2 - \frac{\alpha}{2} \|\mathbf{w}\|^2 \rightarrow \max_w$$

- Решение

$$\mathbf{w}_{MP} = (\beta\Phi^T\Phi + \alpha I)^{-1} \beta\Phi^T \mathbf{t}$$

Линейная регрессия: обсуждение

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- Высокая скорость обучения (достаточно сделать инверсию матрицы $\beta\Phi^T\Phi + \alpha I$ размера $m \times m$)
- Необходимость выбора параметров модели - параметра регуляризации α и обратной дисперсии шума β
- Неразрезанное решение (вообще говоря, все базисные функции входят в решающее правило с ненулевым весом)

Метод релевантных векторов

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- Для получения разреженного решения введем в качестве априорного распределения на параметры \mathbf{w} нормальное распределение с диагональной матрицей ковариации с различными элементами на диагонали:

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

Здесь $A = \text{diag}(\alpha_1, \dots, \alpha_m)$. Такое априорное распределение соответствует независимой регуляризации вдоль каждого веса w_i со своим параметром регуляризации $\alpha_i \geq 0$.

- Для подбора параметров модели $\boldsymbol{\alpha}, \beta$ воспользуемся идеей максимизации обоснованности:

$$p(\mathbf{t}|\boldsymbol{\alpha}, \beta) = \int p(\mathbf{t}|X, \mathbf{w}, \beta)p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \rightarrow \max_{\boldsymbol{\alpha}, \beta}$$

Иллюстрация регуляризации

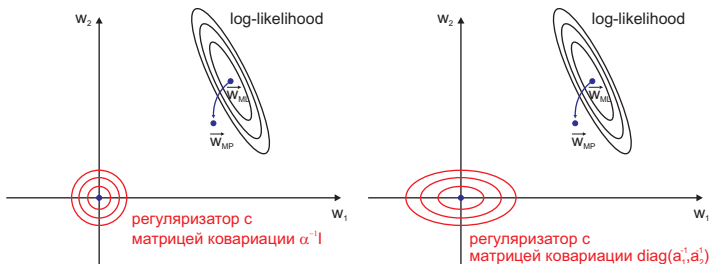
Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации



Вычисление обоснованности

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- Обоснованность является сверткой двух нормальных распределений и может быть вычислена аналитически.

$$p(\mathbf{t}|\boldsymbol{\alpha}, \beta) = \int p(\mathbf{t}|X, \mathbf{w}, \beta)p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} = \int Q(\mathbf{w})d\mathbf{w}$$

- Рассмотрим функцию $L(\mathbf{w}) = \log Q(\mathbf{w})$. Она является квадратичной функцией и может быть представлена как:

$$L(\mathbf{w}) = L(\mathbf{w}_{MP}) + (\nabla_{\mathbf{w}}L(\mathbf{w}_{MP}))^T(\mathbf{w} - \mathbf{w}_{MP}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{MP})^T H(\mathbf{w} - \mathbf{w}_{MP})$$

$$\mathbf{w}_{MP} = \arg \max_{\mathbf{w}} L(\mathbf{w}) \Rightarrow \nabla_{\mathbf{w}}L(\mathbf{w}_{MP}) = 0$$

$$H = \nabla\nabla L(\mathbf{w}_{MP})$$

- Тогда обоснованность может быть вычислена как

$$\begin{aligned} \int Q(\mathbf{w})d\mathbf{w} &= \int \exp\left(L(\mathbf{w}_{MP}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{MP})^T H(\mathbf{w} - \mathbf{w}_{MP})\right) d\mathbf{w} = \\ &= Q(\mathbf{w}_{MP})\sqrt{(2\pi)^m}\sqrt{\det((-H)^{-1})} = \sqrt{(2\pi)^m}\frac{Q(\mathbf{w}_{MP})}{\sqrt{\det(-H)}} \end{aligned}$$

Вычисление обоснованности (Упр.)

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

$$\begin{aligned}L(\mathbf{w}) &= -\frac{\beta}{2}(\mathbf{t} - \Phi\mathbf{w})^T(\mathbf{t} - \Phi\mathbf{w}) - \frac{1}{2}\mathbf{w}^T A \mathbf{w} + C = \\ &= -\frac{\beta}{2}[\mathbf{t}^T \mathbf{t} - 2\mathbf{w}^T \Phi^T \mathbf{t} + \mathbf{w}^T \Phi^T \Phi \mathbf{w}] - \frac{1}{2}\mathbf{w}^T A \mathbf{w} + C\end{aligned}$$

$$\nabla L(\mathbf{w}) = -\frac{1}{2}\beta(-2\Phi^T \mathbf{t} + 2\Phi^T \Phi \mathbf{w}) - A \mathbf{w} = 0 \Rightarrow \mathbf{w}_{MP} = (\beta\Phi^T \Phi + A)^{-1} \beta\Phi^T \mathbf{t}$$

$$\begin{aligned}L(\mathbf{w}_{MP}) &= -\frac{1}{2}[\beta\mathbf{t}^T \mathbf{t} - 2\beta\mathbf{t}^T \Phi(\beta\Phi^T \Phi + A)^{-1} \beta\Phi^T \mathbf{t} + \mathbf{t}^T \Phi \beta(\beta\Phi^T \Phi + A)^{-1} \times \\ &\times (\beta\Phi^T \Phi + A)(\beta\Phi^T \Phi + A)^{-1} \beta\Phi^T \mathbf{t}] + C = -\frac{\beta}{2}\mathbf{t}^T [I - 2\beta\Phi(\beta\Phi^T \Phi + A)^{-1} \Phi^T + \\ &+ \Phi(\beta\Phi^T \Phi + A)^{-1} \beta\Phi^T] \mathbf{t} + C = -\frac{\beta}{2}\mathbf{t}^T [I - \beta\Phi(\beta\Phi^T \Phi + A)^{-1} \Phi^T] \mathbf{t} + C = \\ &= \left\{ (I - \beta\Phi(\beta\Phi^T \Phi + A)^{-1} \Phi^T)^{-1} = \{\text{Тож-во Вудбери}\} = \right. \\ &= \left. I + \beta\Phi(\beta\Phi^T \Phi + A - \Phi^T \beta\Phi)^{-1} \Phi^T = I + \beta\Phi A^{-1} \Phi^T \right\} = \\ &= -0.5\beta\mathbf{t}^T (I + \beta\Phi A^{-1} \Phi^T)^{-1} \mathbf{t} + C = -0.5\mathbf{t}^T (\beta^{-1}I + \Phi A^{-1} \Phi^T)^{-1} \mathbf{t} + C\end{aligned}$$

Вычисление обоснованности (Упр.)

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

$$\begin{aligned} p(\mathbf{t}|X, \boldsymbol{\alpha}, \beta) &= \int p(\mathbf{t}|X, \beta) p(\mathbf{w}|\boldsymbol{\alpha}) d\mathbf{w} = \sqrt{(2\pi)^m} \frac{Q(\mathbf{w}_{MP})}{\sqrt{\det(-H)}} = \\ &= \sqrt{(2\pi)^m} \frac{\sqrt{\beta^n} \sqrt{\det A}}{\sqrt{2\pi^n} \sqrt{(2\pi)^m}} \exp\left(-\frac{1}{2} \mathbf{t}^T (\beta^{-1} I + \Phi A^{-1} \Phi^T)^{-1} \mathbf{t}\right) \frac{1}{\sqrt{\det(-H)}} = \\ &= \left\{ H = -(\beta \Phi^T \Phi + A), \det(-H) = \det(\beta \Phi^T \Phi + A) = \right. \\ &= \left. \{ \text{Лемма об опр-ле матр.} \} = \det(A) \det(I + \beta \Phi A^{-1} \Phi^T) \right\} = \\ &= \frac{1}{\sqrt{(2\pi)^n} \det(\beta^{-1} I + \Phi A^{-1} \Phi^T)^{1/2}} \exp\left(-\frac{1}{2} \mathbf{t}^T (\beta^{-1} I + \Phi A^{-1} \Phi^T)^{-1} \mathbf{t}\right) \end{aligned}$$

Оптимизация обоснованности

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- Приравнивая к нулю производные обоснованности по α, β , можно получить (Упр.) итерационные формулы для пересчета параметров:

$$\alpha_i^{new} = \frac{\gamma_i}{w_{MP,i}^2}$$

$$\gamma_i = 1 - \alpha_i^{old} \Sigma_{ii}$$

$$\beta^{new} = \frac{n - \sum_{i=1}^m \gamma_i}{\|\mathbf{t} - \Phi \mathbf{w}\|^2}$$

Здесь $\Sigma = (\beta \Phi^T \Phi + A)^{-1}$, $\mathbf{w}_{MP} = \beta \Sigma \Phi^T \mathbf{t}$.

- Параметр γ_i может интерпретироваться как степень, в которой соответствующий вес w_i определяется данными или регуляризацией. Если α_i велико, то вес w_i существенно предопределен априорным распределением, $\Sigma_{ii} \simeq \alpha_i^{-1}$ и $\gamma_i \simeq 0$. С другой стороны для малых значений α_i значение веса w_i полностью определяется данными, $\gamma_i \simeq 1$.

Принятие решения

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- Зная значения $\alpha_{MP}, \sigma_{MP}^2$ можно вычислить распределение прогноза (Упр.):

$$p(t_* | \mathbf{x}_*, \mathbf{t}, X) = \int p(t_* | \mathbf{x}_*, \mathbf{w}, \sigma_{MP}^2) p(\mathbf{w} | \mathbf{t}, X, \alpha_{MP}, \sigma_{MP}^2) d\mathbf{w} \sim \mathcal{N}(t_* | y_*, \sigma_*^2)$$

Здесь

$$y_* = \mathbf{w}_{MP}^T \phi(\mathbf{x}_*)$$

$$\sigma_*^2 = \sigma_{MP}^2 + \phi(\mathbf{x}_*)^T \Sigma \phi(\mathbf{x}_*)$$

Метод релевантных векторов для задачи регрессии

Метод релевантных векторов

Ветров, Кропотов

Ликбез

Метод релевантных векторов для задачи регрессии

Метод релевантных векторов для задачи классификации

Вход: Обучающая выборка $\{\mathbf{x}_i, t_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$, $t_i \in \mathbb{R}$;
Матрица обобщенных признаков $\Phi = \{\phi_j(\mathbf{x}_i)\}_{i,j=1}^{n,m}$;

Выход: Набор весов \mathbf{w} , матрица Σ и оценка дисперсии шума β^{-1} для решающего правила $t_*(\mathbf{x}) = \sum_{j=1}^m w_j \phi_j(\mathbf{x})$, $\sigma_*^2(\mathbf{x}) = \beta^{-1} + \Phi^T(\mathbf{x}_*)\Sigma\Phi(\mathbf{x}_*)$;

- 1: инициализация: $\alpha_i := 1$, $i = 1, \dots, m$, $\beta := 1$, AlphaBound := 10^{12} , WeightBound := 10^{-6} , NumberOfIterations := 100;
- 2: **для** $k = 1, \dots, \text{NumberOfIterations}$
- 3: $A := \text{diag}(\alpha_1, \dots, \alpha_m)$;
- 4: $\Sigma := (\beta\Phi^T\Phi + A)^{-1}$;
- 5: $\mathbf{w}_{MP} := \Sigma\beta\Phi^T\mathbf{t}$;
- 6: **для** $j = 1, \dots, m$
- 7: **если** $w_{MP,j} < \text{WeightBound}$ или $\alpha_j > \text{AlphaBound}$ **то**
- 8: $w_{MP,j} := 0$, $\alpha_j := +\infty$, $\gamma_j := 0$;
- 9: **иначе**
- 10: $\gamma_j := 1 - \alpha_j \Sigma_{jj}$, $\alpha_j := \frac{\gamma_j}{w_{MP,j}^2}$;
- 11: $\beta := \frac{n - \sum_{j=1}^m \gamma_j}{\|\mathbf{t} - \Phi\mathbf{w}_{MP}\|^2}$

Метод релевантных векторов. Пример.

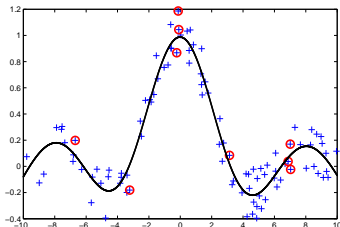
Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации



В процессе обучения большинство α_i стремятся к $+\infty$. Таким образом, априорное распределение на соответствующий вес w_i становится вырожденным, что соответствует ситуации $w_i = 0$, т.е. исключению данного объекта из модели.

Метод релевантных векторов для регрессии: обсуждение

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- На практике процесс обучения обычно требует 20–50 итераций. На каждой итерации вычисляется w_{MP} (это требует обращения матрицы размера $m \times m$), а также пересчитываются значения α , σ^2 (практически не требует времени). Как следствие, скорость обучения метода падает в 20-50 раз по сравнению с линейной регрессией.
- При использовании ядерных функций в качестве обобщенных признаков необходимо проводить скользящий контроль для различных значений параметров ядерных функций. В этом случае время обучения возрастает еще в несколько раз.
- Параметры регуляризации α и дисперсии шума в данных σ^2 подбираются автоматически.
- На выходе получается разреженное решение, т.е. только небольшое количество исходных объектов входят в решающее правило с ненулевым весом.
- Кроме значения прогноза y_* алгоритм выдает также дисперсию прогноза σ_*^2 .

Задача классификации

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- Рассмотрим следующую задачу классификации на два класса: имеется выборка $(X, \mathbf{t}) = \{\mathbf{x}_i, t_i\}_{i=1}^n$, где вектор признаков $\mathbf{x}_i \in \mathbb{R}^d$, а целевая переменная $t_i \in \{+1, -1\}$, требуется для нового объекта \mathbf{x}_* предсказать значение целевой переменной t_* .
- Воспользуемся обобщенными линейными моделями для классификации:

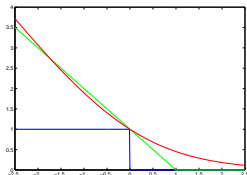
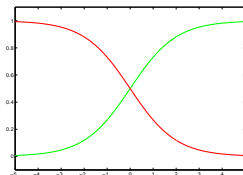
$$\hat{t}(\mathbf{x}) = \text{sign}(f(\mathbf{x})) = \text{sign} \left(\sum_{j=1}^m w_j \phi_j(\mathbf{x}) \right) = \text{sign}(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}))$$

Здесь \mathbf{w} — набор числовых параметров, а $\boldsymbol{\phi}(\mathbf{x})$ — вектор обобщенных признаков.

Метод максимума правдоподобия (логистическая регрессия)

- В качестве функции правдоподобия выберем произведение логистических функций:

$$p(\mathbf{t}|X, \mathbf{w}) = \prod_{i=1}^n p(t_i|\mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^n \frac{1}{1 + \exp(-t_i f(\mathbf{x}_i))}$$



- Переходя к логарифму правдоподобия, получаем:

$$-\sum_{i=1}^n \log(1 + \exp(-t_i f(\mathbf{x}_i))) \rightarrow \max_{\mathbf{w}}$$

Оптимизация функции правдоподобия (IRLS)

Метод релевантных векторов

Ветров, Кропотов

Ликбез

Метод релевантных векторов для задачи регрессии

Метод релевантных векторов для задачи классификации

- Функция $-\log(1 + \exp(-x))$ является вогнутой, поэтому логарифм правдоподобия как сумма вогнутых функций также является вогнутой функцией и имеет единственный максимум.
- Для поиска максимума логарифма правдоподобия L воспользуемся методом Ньютона:

$$\mathbf{w}^{new} = \mathbf{w}^{old} - H^{-1} \nabla L(\mathbf{w})$$

Здесь $H = \nabla \nabla L(\mathbf{w})$ — гессиан логарифма правдоподобия.

- Вычисляя градиент и гессиан, получаем формулы пересчета:

$$\begin{aligned} \mathbf{w}^{new} &= (\Phi^T R \Phi)^{-1} \Phi^T R \mathbf{z} \\ \mathbf{z} &= \Phi \mathbf{w}^{old} + R^{-1} \text{diag}(\mathbf{t}) \mathbf{s} \end{aligned}$$

Здесь

$$s_i = \frac{1}{1 + \exp(t_i f(x_i))}, \quad R = \text{diag}(s_1(1 - s_1), \dots, s_n(1 - s_n)).$$

Введение регуляризации

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- По аналогии с линейной регрессией можно рассмотреть максимум апостериорной плотности с нормальным априорным распределением с единичной матрицей ковариации, умноженной на коэффициент α^{-1} :

$$-\sum_{i=1}^n \log(1 + \exp(-t_i f(\mathbf{x}_i))) - \frac{\alpha}{2} \|\mathbf{w}\|^2 \rightarrow \max_{\mathbf{w}}$$

- Метод оптимизации меняется следующим образом:

$$\mathbf{w}^{new} = (\Phi^T R \Phi + \alpha I)^{-1} \Phi^T R \mathbf{z}$$

$$\mathbf{z} = \Phi \mathbf{w}^{old} + R^{-1} \text{diag}(\mathbf{t}) \mathbf{s}$$

Логистическая регрессия: обсуждение

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- По-прежнему довольно высокая скорость работы. На практике обучение часто требует всего 3-7 итераций.
- Необходимость выбора параметра регуляризации α
- Неразрезанное решение

Метод релевантных векторов

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- Для получения разреженного решения введем в качестве априорного распределения на параметры \mathbf{w} нормальное распределение с диагональной матрицей ковариации с различными элементами на диагонали:

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

Здесь $A = \text{diag}(\alpha_1, \dots, \alpha_m)$.

- Для подбора параметров модели $\boldsymbol{\alpha}$ воспользуемся идеей максимизации обоснованности:

$$p(\mathbf{t}|\boldsymbol{\alpha}) = \int p(\mathbf{t}|X, \mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \rightarrow \max_{\boldsymbol{\alpha}}$$

Приближение Лапласа

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

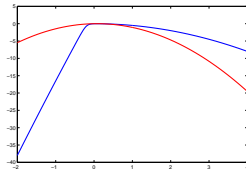
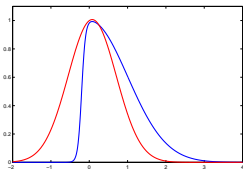
Метод
релевантных
векторов для
задачи
классификации

- 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:

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



Вычисление обоснованности

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- Подынтегральное выражение в обоснованности является произведением логистических функций и нормального распределения. Такой интеграл не берется аналитически.
- Решение: приблизить подынтегральную функцию гауссианой, интеграл от которой легко берется. Для приближения воспользуемся методом Лапласа:

$$p(\mathbf{t}|\boldsymbol{\alpha}) = \int p(\mathbf{t}|X, \mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} = \int Q(\mathbf{w})d\mathbf{w} \simeq \\ \simeq \sqrt{(2\pi)^m} \frac{Q(\mathbf{w}_{MP})}{\sqrt{\det(-\nabla\nabla \log Q(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_{MP}})}}$$

$$\mathbf{w}_{MP} = \arg \max_{\mathbf{w}} Q(\mathbf{w})$$

Оптимизация обоснованности

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

Приравнивая к нулю производную логарифма обоснованности по α , получаем:

$$\begin{aligned} \log p(t|X, \alpha) &= \log Q(\mathbf{w}_{MP}) - \frac{1}{2} \log \det(-H) + C = \\ &= - \sum_{i=1}^n \log(1 + \exp(-t_i f(\mathbf{x}_i, \mathbf{w}_{MP}))) - \frac{1}{2} \sum_{i=1}^M \alpha_i w_{MP,i}^2 - \frac{1}{2} \log \det(-H) + \frac{1}{2} \sum_{i=1}^m \log \alpha_i + C \end{aligned}$$

Здесь $H = -\Phi^T R \Phi - A$, $R = \text{diag}(s_1(1 - s_1), \dots, s_n(1 - s_n))$, $s_i = \frac{1}{1 + \exp(t_i f(\mathbf{x}_i, \mathbf{w}_{MP}))}$.

$$\frac{\partial}{\partial \alpha_j} \log p(t|X, \alpha) = (\text{Упр.}) = -\frac{1}{2} w_{MP,j}^2 - \frac{1}{2} ((\Phi^T R \Phi + A)^{-1})_{jj} + \frac{1}{2\alpha_j} = 0$$

Отсюда получаем итерационные формулы пересчета α , аналогичные регрессии:

$$\alpha_i^{new} = \frac{1 - \alpha_i^{old} \sum_{ii}}{w_{MP,i}^2}$$

Метод релевантных векторов для задачи классификации

Метод релевантных векторов

Ветров, Кропотов

Ликбез

Метод релевантных векторов для задачи регрессии

Метод релевантных векторов для задачи классификации

Вход: Обучающая выборка $\{\mathbf{x}_i, t_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$, $t_i \in \{+1, -1\}$;

Матрица обобщенных признаков $\Phi = \{\phi_j(\mathbf{x}_i)\}_{i,j=1}^{n,m}$;

Выход: Набор весов \mathbf{w} для решающего правила $t_*(\mathbf{x}) = \sum_{j=1}^m w_j \phi_j(\mathbf{x})$;

1: инициализация: $\alpha_i := 1$, $i = 1, \dots, m$, $\mathbf{w}_{MP} = \mathbf{t}$, AlphaBound := 10^{12} , WeightBound := 10^{-6} , NumberOfIterations := 100;

2: **для** $k = 1, \dots, \text{NumberOfIterations}$

3: $A := \text{diag}(\alpha_1, \dots, \alpha_m)$;

4: **повторять**

5: **для** $i = 1, \dots, n$

6: $s_i := 1 / (1 + \exp(t_i \sum_{j=1}^m w_{MP,j} \phi_j(\mathbf{x}_i)))$;

7: $R := \text{diag}(s_1(1 - s_1), \dots, s_n(1 - s_n))$;

8: $\mathbf{z} := \Phi \mathbf{w}_{MP} + R^{-1}(\mathbf{s} - \mathbf{t})$;

9: $\Sigma := (\Phi^T R \Phi + A)^{-1}$;

10: $\mathbf{w}_{MP} := \Sigma \Phi^T R \mathbf{z}$;

11: **пока** $\|\mathbf{w}_{MP}^{new} - \mathbf{w}_{MP}^{old}\|$ меняется больше, чем на заданную величину

12: **для** $j = 1, \dots, m$

13: **если** $w_{MP,j} < \text{WeightBound}$ или $\alpha_j > \text{AlphaBound}$ **то**

14: $w_{MP,j} := 0$, $\alpha_j := +\infty$, $\gamma_j := 0$;

15: **иначе**

16: $\alpha_j := \frac{1 - \alpha_j \Sigma_{jj}}{w_{MP,j}^2}$;

Метод релевантных векторов. Пример.

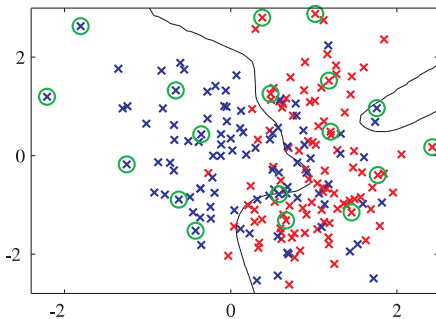
Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации



В процессе обучения большинство α_i стремятся к $+\infty$. Таким образом, априорное распределение на соответствующий вес w_i становится вырожденным, что соответствует ситуации $w_i = 0$, т.е. исключению данной базисной функции из модели.

Метод релевантных векторов: обсуждение

Метод
релевантных
векторов

Ветров,
Кропотов

Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации

- На практике процесс обучения обычно требует 20–50 итераций. На каждой итерации вычисляется w_{MP} (это требует 3–7 итераций с обращениями матрицы размера $m \times m$), а также пересчитываются значения α (практически не требует времени). Как следствие, скорость обучения метода падает в 20–50 раз по сравнению с логистической регрессией.
- При использовании ядерных функций в качестве обобщенных признаков необходимо проводить скользящий контроль для различных значений параметров ядерных функций. В этом случае время обучения возрастает еще в несколько раз.
- Параметры регуляризации α подбираются автоматически.
- На выходе получается разреженное решение.
- Для вычисления дисперсии прогноза необходимо проводить дополнительно аппроксимацию интеграла

$$p(t_* | \mathbf{x}_*, t, X) = \int p(t_* | \mathbf{x}_*, \mathbf{w}) p(\mathbf{w} | t, X, \alpha_{MP}) d\mathbf{w}$$

RVM vs. SVM

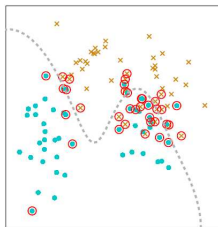
Метод
релевантных
векторов

Ветров,
Кропотов

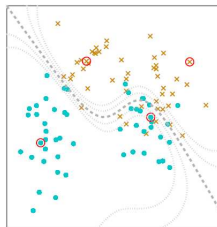
Ликбез

Метод
релевантных
векторов для
задачи регрессии

Метод
релевантных
векторов для
задачи
классификации



SVM



RVM

- SVM:

$$\sum_{i=1}^n [1 - t_i f(\mathbf{x}_i, \mathbf{w})]_+ + \gamma \sum_{j=1}^m w_j^2 \rightarrow \min_{\mathbf{w}}$$

- RVM:

$$\sum_{i=1}^n \log(1 + \exp(-t_i f(\mathbf{x}_i, \mathbf{w}))) + \sum_{j=1}^m \alpha_j w_j^2 \rightarrow \min_{\mathbf{w}}$$