Лекция 6. Метод релевантных векторов

Д.П. Ветров[1] Д.А. Кропотов[2]

[1]МГУ, ВМиК, каф. ММП

[2]ВЦ РАН

Спецкурс «Байесовские методы машинного обучения»

План лекции

Ликбез

Метод релевантных векторов для задачи регрессии

Метод релевантных векторов для задачи классификации

Матричные тождества

- Тождество 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

- Тождество

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

- Доказательство

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

Тождества для определителей матрицы

- 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)

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

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

  Здесь $\boldsymbol{w}$ — набор числовых параметров, а $\boldsymbol{\phi}(\boldsymbol{x})$ — вектор обобщенных признаков.
- Часто в качестве обобщенных признаков выбираются следующие:
  - Обычные признаки — $\phi_j(\boldsymbol{x}) = x_j,\ j = 1, \ldots, d$
  - Ядровые функции — $\phi_j(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{x}_j),\ j = 1, \ldots, n,\ \phi_{n+1}(\boldsymbol{x}) \equiv 1$

Метод максимума правдоподобия (линейная регрессия)

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

$$p(\boldsymbol{t}|X,\boldsymbol{w}) = \prod_{i=1}^{n} p(t_i|\boldsymbol{x}_i,\boldsymbol{w}) = \prod_{i=1}^{n} \mathcal{N}(t_i|f(\boldsymbol{x}_i,\boldsymbol{w}),\beta^{-1}) =$$

$$\prod_{i=1}^{n} \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}(t_i - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2\right) = \sqrt{\frac{\beta}{2\pi}}^n \exp\left(-\frac{\beta}{2}\sum_{i=1}^{n}(t_i - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2\right)$$

- Переходя к логарифму, получаем

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

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

- Точка максимума правдоподобия выписывается в явном виде:

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

Введение регуляризации (априорного распределения)

- Следуя байесовскому подходу воспользуемся методом максимума апостериорной плотности:

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

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

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

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

- Максимизация апостериорной плотности эквивалентна следующей задаче оптимизации:

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

- Решение

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

Линейная регрессия: обсуждение

- 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
- Неразреженное решение (вообще говоря, все базисные функции входят в решающее правило с ненулевым весом)

Метод релевантных векторов

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

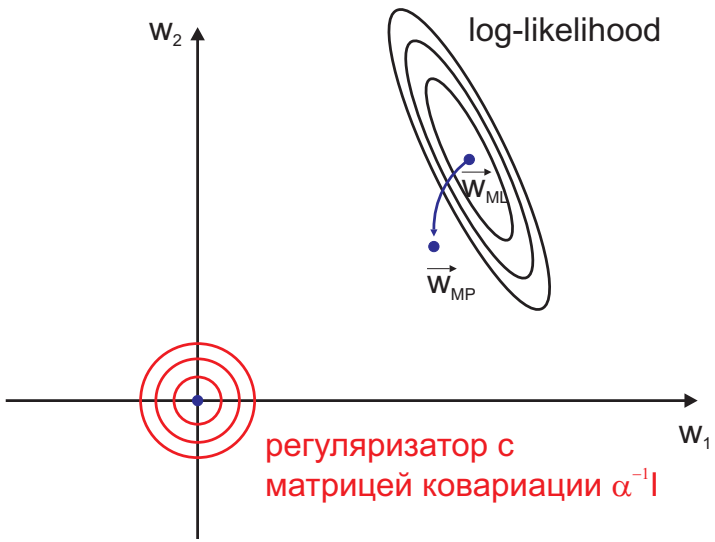
$$p(\boldsymbol{w}|\boldsymbol{\alpha}) \sim \mathcal{N}(0, A^{-1})$$

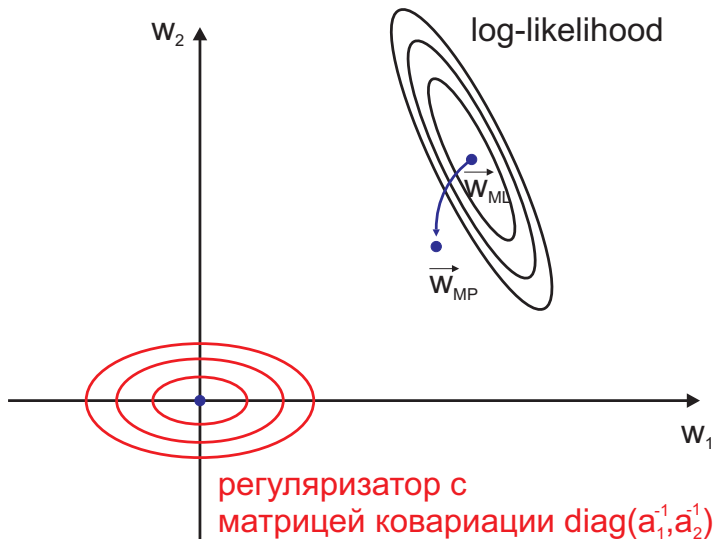
Здесь $A = \text{diag}(\alpha_1, \ldots, \alpha_m)$. Такое априорное распределение соответствует независимой регуляризации вдоль каждого веса w_i со своим параметром регуляризации $\alpha_i \geq 0$.

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

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

Иллюстрация регуляризации





Вычисление обоснованности

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

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

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

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

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

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

- Тогда обоснованность может быть вычислена как

$$\int Q(\boldsymbol{w})d\boldsymbol{w} = \int \exp\left(L(\boldsymbol{w}_{MP}) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{MP})^T H(\boldsymbol{w} - \boldsymbol{w}_{MP})\right) d\boldsymbol{w} =$$

$$= Q(\boldsymbol{w}_{MP})\sqrt{(2\pi)^m}\sqrt{\det((-H)^{-1})} = \sqrt{(2\pi)^m}\frac{Q(\boldsymbol{w}_{MP})}{\sqrt{\det(-H)}}$$

Вычисление обоснованности (Упр.)

$$L(\boldsymbol{w}) = -\frac{\beta}{2}(\boldsymbol{t} - \Phi\boldsymbol{w})^T(\boldsymbol{t} - \Phi\boldsymbol{w}) - \frac{1}{2}\boldsymbol{w}^T A\boldsymbol{w} + C =$$

$$= -\frac{\beta}{2}[\boldsymbol{t}^T\boldsymbol{t} - 2\boldsymbol{w}^T\Phi^T\boldsymbol{t} + \boldsymbol{w}^T\Phi^T\Phi\boldsymbol{w}] - \frac{1}{2}\boldsymbol{w}^T A\boldsymbol{w} + C$$

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

$$L(\boldsymbol{w}_{MP}) = -\frac{1}{2}[\beta\boldsymbol{t}^T\boldsymbol{t} - 2\beta\boldsymbol{t}^T\Phi(\beta\Phi^T\Phi + A)^{-1}\beta\Phi^T\boldsymbol{t} + \boldsymbol{t}^T\Phi\beta(\beta\Phi^T\Phi + A)^{-1}\times$$

$$\times\cancel{(\beta\Phi^T\Phi + A)}\;\cancel{(\beta\Phi^T\Phi + A)^{-1}}\;\beta\Phi^T\boldsymbol{t}] + C = -\frac{\beta}{2}\boldsymbol{t}^T[I - 2\beta\Phi(\beta\Phi^T\Phi + A)^{-1}\Phi^T +$$

$$+ \Phi(\beta\Phi^T\Phi + A)^{-1}\beta\Phi^T]\boldsymbol{t} + C = -\frac{\beta}{2}\boldsymbol{t}^T[I - \beta\Phi(\beta\Phi^T\Phi + A)^{-1}\Phi^T]\boldsymbol{t} + C =$$

$$\left\{(I - \beta\Phi(\beta\Phi^T\Phi + A)^{-1}\Phi^T)^{-1} = \{\text{Тож-во Вудбери}\} =\right.$$

$$\left. = I + \beta\Phi(\cancel{\beta\Phi^T\Phi} + A - \cancel{\Phi^T\beta\Phi})^{-1}\Phi^T = I + \beta\Phi A^{-1}\Phi^T\right\} =$$

$$= -0.5\beta\boldsymbol{t}^T(I + \beta\Phi A^{-1}\Phi^T)^{-1}\boldsymbol{t} + C = -0.5\boldsymbol{t}^T(\beta^{-1}I + \Phi A^{-1}\Phi^T)^{-1}\boldsymbol{t} + C$$

Вычисление обоснованности (Упр.)

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

$$= \sqrt{(2\pi)^m}\frac{\sqrt{\beta^n}\sqrt{\det A}}{\sqrt{2\pi}^n\sqrt{(2\pi)^m}}\exp\left(-\frac{1}{2}\boldsymbol{t}^T(\beta^{-1}I+\Phi A^{-1}\Phi^T)^{-1}\boldsymbol{t}\right)\frac{1}{\sqrt{\det(-H)}} =$$

$$= \Big\{H = -(\beta\Phi^T\Phi + A),\ \det(-H) = \det(\beta\Phi^T\Phi + A) =$$

$$= \{\text{Лемма об опр-ле матр.}\} = \det(A)\det(I+\beta\Phi A^{-1}\Phi^T)\Big\} =$$

$$= \frac{1}{\sqrt{(2\pi)^n}\det(\beta^{-1}I+\Phi A^{-1}\Phi^T)^{1/2}}\exp\left(-\frac{1}{2}\boldsymbol{t}^T(\beta^{-1}I+\Phi A^{-1}\Phi^T)^{-1}\boldsymbol{t}\right)$$

Оптимизация обоснованности

- Приравнивая к нулю производные обоснованности по $\boldsymbol{\alpha}, \beta$, можно получить (Упр.) итерационные формулы для пересчета параметров:

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

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

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

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

- 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

- Зная значения $\boldsymbol{\alpha}_{MP}, \sigma^2_{MP}$ можно вычислить распределение прогноза (Упр.):

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

Здесь

$$y_* = \boldsymbol{w}^T_{MP} \boldsymbol{\phi}(\boldsymbol{x}_*)$$
$$\sigma^2_* = \sigma^2_{MP} + \boldsymbol{\phi}(\boldsymbol{x}_*)^T \Sigma \boldsymbol{\phi}(\boldsymbol{x}_*)$$

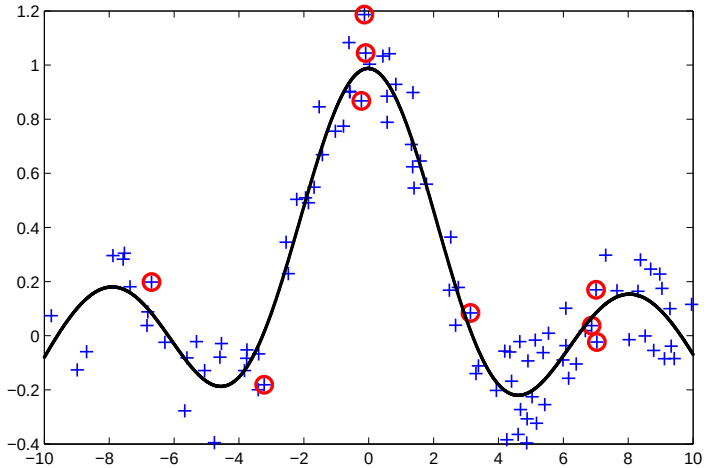
Метод релевантных векторов для задачи регрессии

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

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

1: инициализация: $\alpha_i := 1$, $i = 1, \ldots, m$, $\beta := 1$, AlphaBound := 10^{12}, WeightBound := 10^{-6}, NumberOfIterations := 100;
2: для $k = 1, \ldots,$ NumberOfIterations
3: $A := \mathrm{diag}(\alpha_1, \ldots, \alpha_m)$;
4: $\Sigma := (\beta\Phi^T\Phi + A)^{-1}$;
5: $\boldsymbol{w}_{MP} := \Sigma\beta\Phi^T\boldsymbol{t}$;
6: для $j = 1, \ldots, m$
7: если $w_{MP,j} <$ WeightBound или $\alpha_j >$ 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: $\beta := \frac{n - \sum_{j=1}^m \gamma_j}{\|\boldsymbol{t} - \Phi\boldsymbol{w}_{MP}\|^2}$

Метод релевантных векторов. Пример.



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

- На практике процесс обучения обычно требует 20–50 итераций. На каждой итерации вычисляется $\boldsymbol{w}_{MP}$ (это требует обращения матрицы размера $m \times m$), а также пересчитываются значения $\boldsymbol{\alpha}$, σ^2 (практически не требует времени). Как следствие, скорость обучения метода падает в 20-50 раз по сравнению с линейной регрессией.
- При использовании ядровых функций в качестве обобщенных признаков необходимо проводить скользящий контроль для различных значений параметров ядровых функций. В этом случае время обучения возрастает еще в несколько раз.
- Параметры регуляризации $\boldsymbol{\alpha}$ 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

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

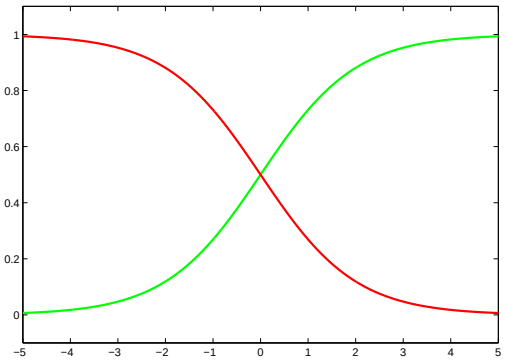
$$\hat{t}(\boldsymbol{x}) = \text{sign}(f(\boldsymbol{x})) = \text{sign}\left(\sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x})\right) = \text{sign}\left(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})\right)$$

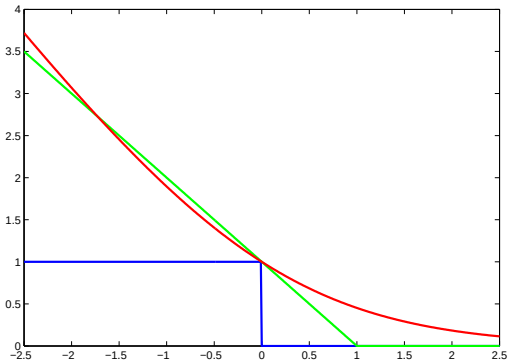
Здесь $\boldsymbol{w}$ — набор числовых параметров, а $\boldsymbol{\phi}(\boldsymbol{x})$ — вектор обобщенных признаков.

Метод максимума правдоподобия (логистическая регрессия)

- В качестве функции правдоподобия выберем произведение логистических функций:

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





- Переходя к логарифму правдоподобия, получаем:

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

Оптимизация функции правдоподобия (IRLS)

- 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:

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

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

- Вычисляя градиент и гессиан, получаем формулы пересчета:

$$\boldsymbol{w}^{new} = (\Phi^T R\Phi)^{-1}\Phi^T R\boldsymbol{z}$$

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

Здесь

$s_i = \frac{1}{1+\exp(t_i f(\boldsymbol{x}_i))}$, $\quad R = \operatorname{diag}(s_1(1-s_1),\ldots,s_n(1-s_n))$.

Введение регуляризации

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

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

- Метод оптимизации меняется следующим образом:

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

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

Логистическая регрессия: обсуждение

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

Метод релевантных векторов

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

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

  Здесь $A = \text{diag}(\alpha_1, \ldots, \alpha_m)$.
- Для подбора параметров модели $\boldsymbol{\alpha}$ воспользуемся идеей максимизации обоснованности:

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

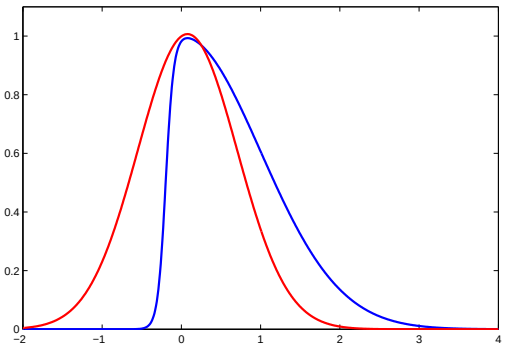
Приближение Лапласа

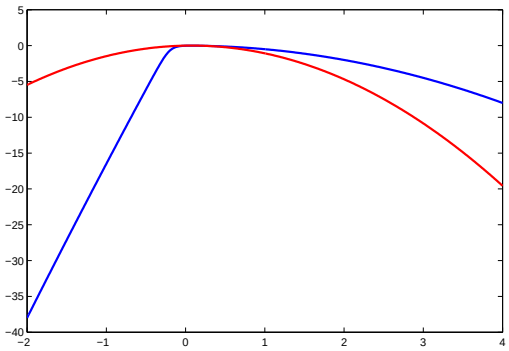
- Рассмотрим функцию $p(z) = \exp\left(-\frac{z^2}{2}\right)\frac{1}{1+\exp(-20z-4)}$.
- Разложим логарифм функции в ряд Тейлора в точке максимума:

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

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

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





Вычисление обоснованности

- 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:

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

$$\simeq \sqrt{(2\pi)^m}\frac{Q(\boldsymbol{w}_{MP})}{\sqrt{\det(-\nabla\nabla\log Q(\boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}_{MP}})}}$$

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

Оптимизация обоснованности

Приравнивая к нулю производную логарифма обоснованности по $\boldsymbol{\alpha}$, получаем:

$$\log p(\boldsymbol{t}|X,\boldsymbol{\alpha}) = \log Q(\boldsymbol{w}_{MP}) - \frac{1}{2}\log\det(-H) + C =$$

$$-\sum_{i=1}^{n}\log(1+\exp(-t_i f(\boldsymbol{x}_i,\boldsymbol{w}_{MP}))) - \frac{1}{2}\sum_{i=1}^{M}\alpha_i w_{MP,i}^2 - \frac{1}{2}\log\det(-H) + \frac{1}{2}\sum_{i=1}^{m}\log\alpha_i + C$$

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

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

Отсюда получаем итерационные формулы пересчета $\boldsymbol{\alpha}$, аналогичные регрессии:

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

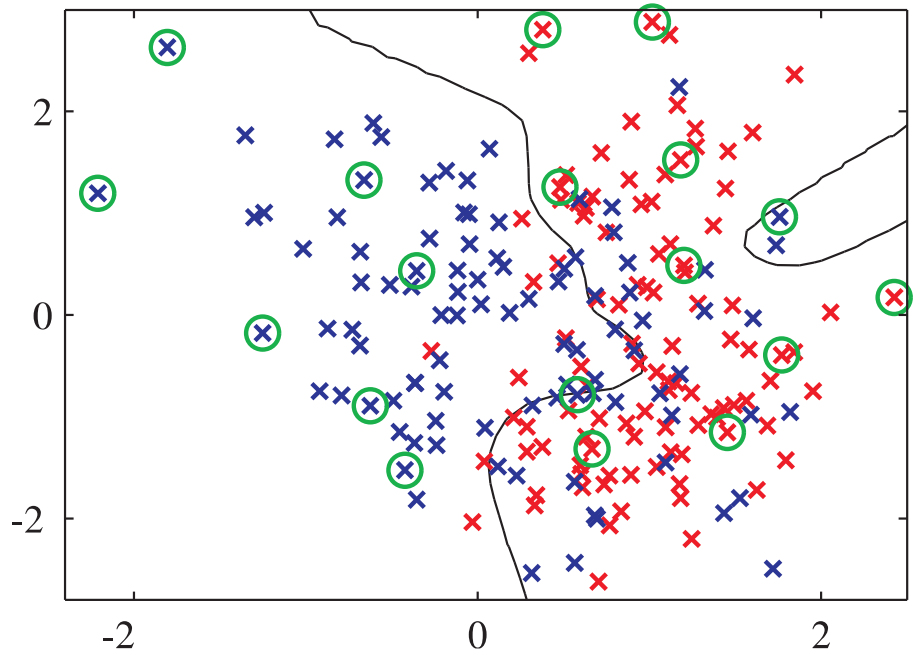
Метод релевантных векторов для задачи классификации

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

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

1: инициализация: $\alpha_i := 1$, $i = 1, \ldots, m$, $\boldsymbol{w}_{MP} = \boldsymbol{t}$, AlphaBound := 10^{12}, WeightBound := 10^{-6}, NumberOfIterations := 100;
2: для $k = 1, \ldots,$ NumberOfIterations
3: $A := \mathrm{diag}(\alpha_1, \ldots, \alpha_M)$;
4: повторять
5: для $i = 1, \ldots, n$
6: $s_i := 1/(1 + \exp(t_i \sum_{j=1}^m w_{MP,j} \phi_j(\boldsymbol{x}_i)))$;
7: $R := \mathrm{diag}(s_1(1 - s_1), \ldots, s_n(1 - s_n))$;
8: $\boldsymbol{z} := \Phi \boldsymbol{w}_{MP} + R^{-1}(\boldsymbol{s} - \boldsymbol{t})$;
9: $\Sigma := (\Phi^T R \Phi + A)^{-1}$;
10: $\boldsymbol{w}_{MP} := \Sigma \Phi^T R \boldsymbol{z}$;
11: пока $\|\boldsymbol{w}_{MP}^{new} - \boldsymbol{w}_{MP}^{old}\|$ меняется больше, чем на заданную величину
12: для $j = 1, \ldots, m$
13: если $w_{MP,j} <$ WeightBound или $\alpha_j >$ 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.



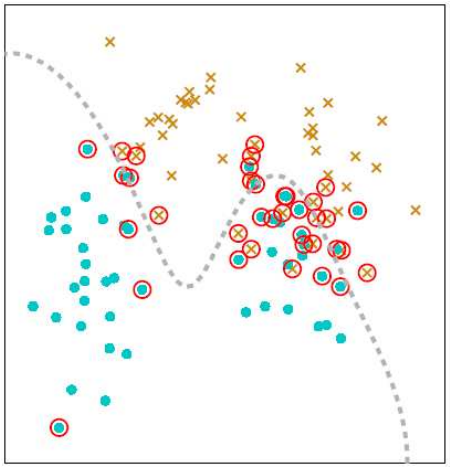
В процессе обучения большинство α_i стремятся к $+\infty$. Таким образом, априорное распределение на соответствующий вес w_i становится вырожденным, что соответствует ситуации $w_i = 0$, т.е. исключению данной базисной функции из модели.

Метод релевантных векторов: обсуждение

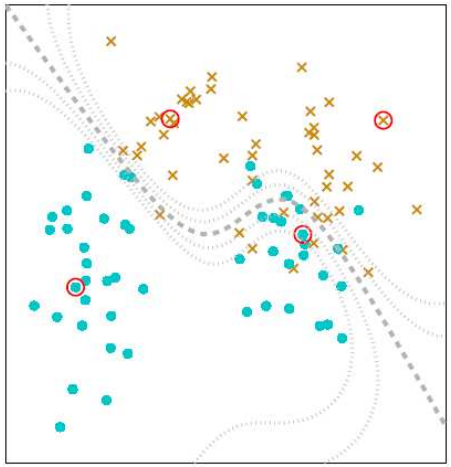
- На практике процесс обучения обычно требует 20–50 итераций. На каждой итерации вычисляется $\boldsymbol{w}_{MP}$ (это требует 3–7 итераций с обращениями матрицы размера $m \times m$), а также пересчитываются значения $\boldsymbol{\alpha}$ (практически не требует времени). Как следствие, скорость обучения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

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

RVM vs. SVM



SVM



RVM

- SVM:

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

- RVM:

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