

Машинное обучение: Линейный модели классификации и регрессии

MADE academy

Эмили Драль

Алгоритмы машинного обучения

1. **Обучение с учителем: линейные модели**
2. Обучение с учителем: ансамбли моделей
3. Обучение с учителем: нейросетевые модели
4. Обучение без учителя: обзор методов
5. (optional) Рекомендательные системы
6. (optional) Обучение с подкреплением

План занятия

1. Линейные модели: интуиция
2. Формализация линейной модели
3. Построение линейной модели
4. Часто используемые линейные модели

Линейные модели: интуиция

Линейные
модели:
интуиция

Линейная модель



Стоит ли занимать очередь в банке?

Пример:
занимать ли
очередь в
банке?

Признаки (1/0)



Вы свободны в
данный момент



Вы голодны



Вам хочется спать



В очереди меньше 3х
человек

Пример:
занимать ли
очередь в
банке?

Признаки (1/0)



Вы свободны в
данный момент



Вы голодны



Вам хочется спать



В очереди меньше 3х
человек

Пример:
занимать ли
очередь в
банке?

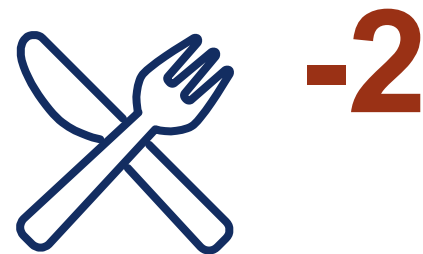
Признаки (1/0)



Вы свободны в
данный момент



Вам хочется спать



Вы голодны



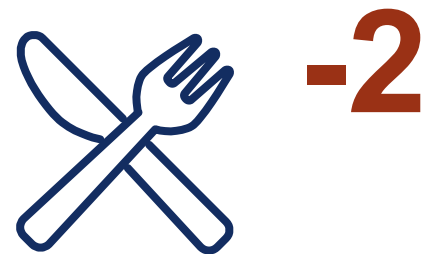
В очереди меньше 3х
человек

Пример:
занимать ли
очередь в
банке?

Признаки (1/0)



Вы свободны в
данный момент



Вы голодны



Вам хочется спать



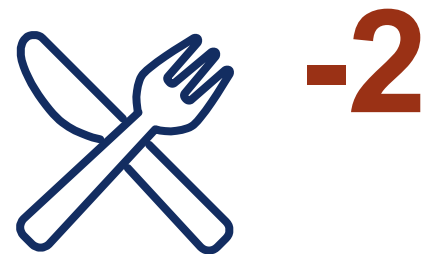
В очереди меньше 3х
человек

Пример:
занимать ли
очередь в
банке?

Признаки (1/0)



Вы свободны в
данный момент



Вы голодны



Вам хочется спать



В очереди меньше 3х
человек

Линейные модели

Пример:
занимать ли
очередь в
банке?



Порог для решающего правила: 0
Если сумма больше 0 – занимаем очередь!

Линейные модели: интуиция

Формализуем модель

$$a(x) = \begin{cases} 1, & \text{если } f(x) \geq 0 \\ -1, & \text{если } f(x) < 0 \end{cases}$$

$$f(\mathbf{x}) = w_0 + w_1x_1 + \dots + w_nx_n$$

Линейные модели: интуиция

Формализуем модель

$$a(x) = \begin{cases} 1, & \text{если } f(x) \geq 0 \\ -1, & \text{если } f(x) < 0 \end{cases}$$

$$f(x) = w_0 + w_1x_1 + \dots + w_nx_n = \langle w, x \rangle + w_0$$

Введём фиктивный признак x_0 , равный 1 на всех объектах

$$f(x) = \langle w, x \rangle$$

Пример:
скоринговые
карты

Показатель	Диапазон значений
Возраст заёмщика	До 35 лет
	От 35 до 45 лет
	От 45 и старше
Образование	Высшее
	Среднее специальное
	Среднее
Состоит ли в браке	Да
	Нет
Наличие кредита в прошлом	Да
	Нет
Стаж работы	До 1 года
	От 1 до 3 лет
	От 3 до 6 лет
	Свыше 6 лет
Наличие автомобиля	Да
	Нет

Пример:
скоринговые
карты

Показатель	Диапазон значений	Скоринг-балл
Возраст заёмщика	До 35 лет	7,60
	От 35 до 45 лет	29,68
	От 45 и старше	15,87
Образование	Высшее	29,82
	Среднее специальное	20,85
	Среднее	22,71
Состоит ли в браке	Да	29,46
	Нет	9,38
Наличие кредита в прошлом	Да	40,55
	Нет	13,91
Стаж работы	До 1 года	15,00
	От 1 до 3 лет	18,14
	От 3 до 6 лет	19,85
	Свыше 6 лет	23,74
Наличие автомобиля	Да	51,69
	Нет	15,93

Пример:
скоринговые
карты

Почему нельзя продолжать так же?

- Сложно настраивать вручную
- Требуется эксперт в области
- Требуется проверка на данных, уточнение (калибровка) весов

Пример:
скоринговые
карты

Почему нельзя продолжать так же?

- Сложно настраивать вручную
- Требуется эксперт в области
- Требуется проверка на данных, уточнение (калибровка) весов

Решение

Автоматизируем подбор параметров:
придумаем функцию от параметров, которую
надо минимизировать, и используем методы
численной оптимизации

Пример: скоринговые карты

Почему нельзя продолжать так же?

- Сложно настраивать вручную
- Требуется эксперт в области
- Требуется проверка на данных, уточнение (калибровка) весов

Решение

Автоматизируем подбор параметров:
придумаем функцию от параметров – $Q(a)$,
которую надо минимизировать $Q(a) \rightarrow \min$, и
используем методы численной оптимизации

Формализация линейной модели

Линейная регрессия

$$a(x) = \langle w, x \rangle + w_0$$

Линейные
модели:
формализация

Линейные
модели:
формализация

Линейная регрессия

$$a(x) = \langle w, x \rangle + w_0$$

А как получить ответ в задаче классификации?

Линейная классификация

$$a(x) = \langle w, x \rangle + w_0$$

А как получить ответ в задаче классификации?

- Выберем метки класса 1 и -1 для удобства
- Формализуем пороговое решающее правило

$$a(x) = \text{sign}(\langle w, x \rangle + w_0)$$

Если скалярное произведение неотрицательное – класс 1, в противном случае класс -1

Линейные модели: формализация

Интерпретация

$$a(x) = \langle w, x \rangle + w_0$$

- Абсолютные значения весов w_1, \dots, w_n — можно интерпретировать как важность признаков
- Знак можно интерпретировать как класс, за который "голосует" признак

*для данной интерпретации признаки должны быть откалиброваны или бинаризованы, об этом позднее

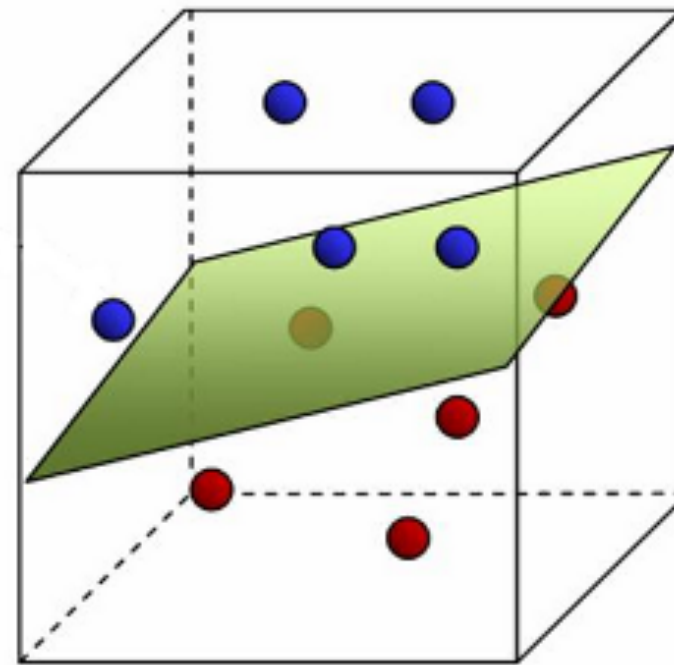
Линейные модели: формализация

Геометрическая интерпретация

$$a(x) = \langle w, x \rangle + w_0$$

Веса задают гиперплоскость, разделяющую классы

$$a(x) = \begin{cases} 1, & \text{если } f(x) > 0 \\ -1, & \text{если } f(x) \leq 0 \end{cases}$$

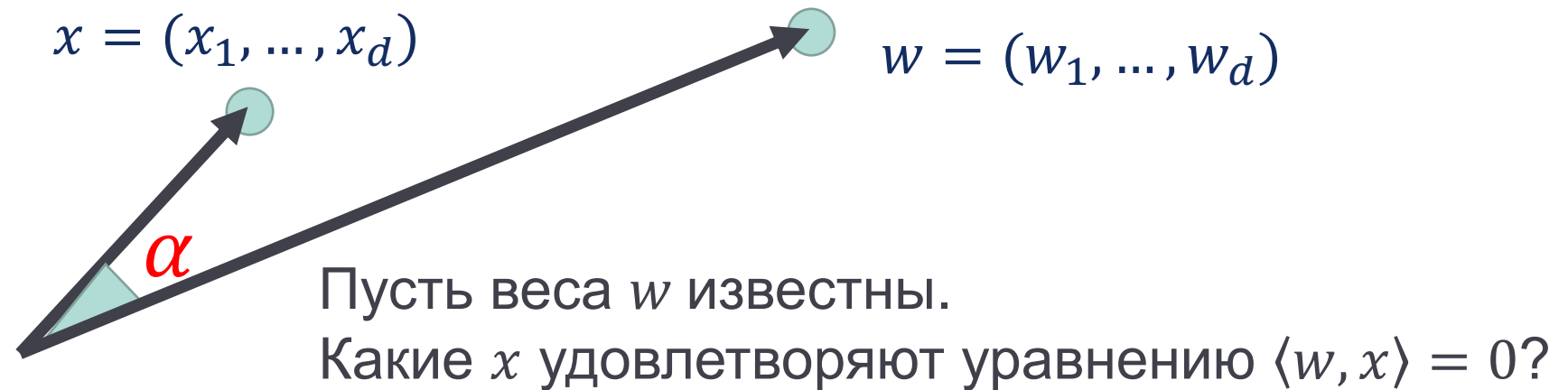


Линейные модели: формализация

Геометрическая интерпретация

$$a(x) = \langle w, x \rangle + w_0$$

Веса задают гиперплоскость, разделяющую классы.
Откуда берется гиперплоскость?

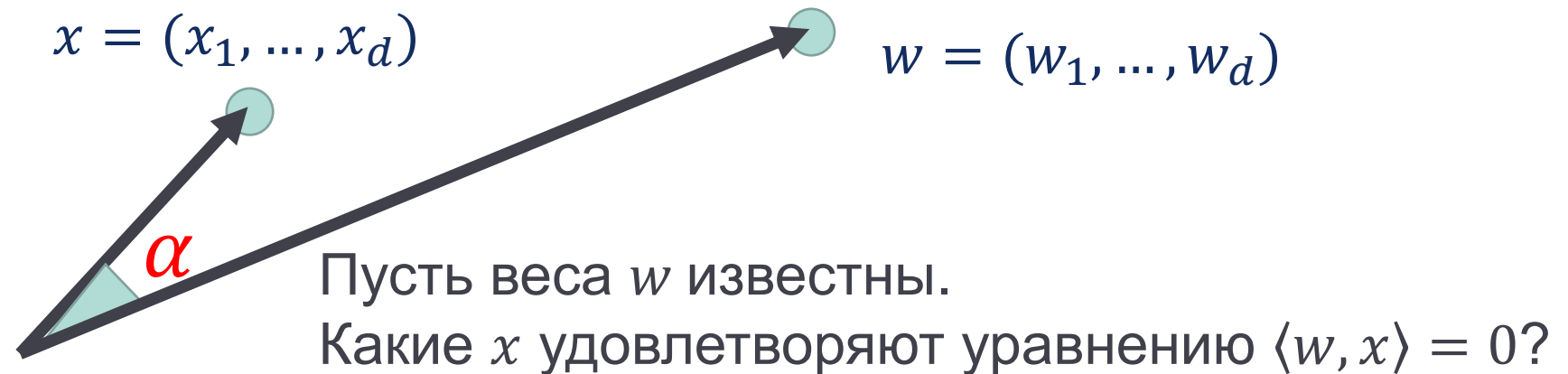


Линейные модели: формализация

Геометрическая интерпретация

$$a(x) = \langle w, x \rangle + w_0$$

Веса задают гиперплоскость, разделяющую классы.
Откуда берется гиперплоскость?



$$\langle w, x \rangle = \|w\| \|x\| \cos \alpha = 0,$$

Значит, если оба вектора не нулевые:

$$\cos \alpha = 0$$

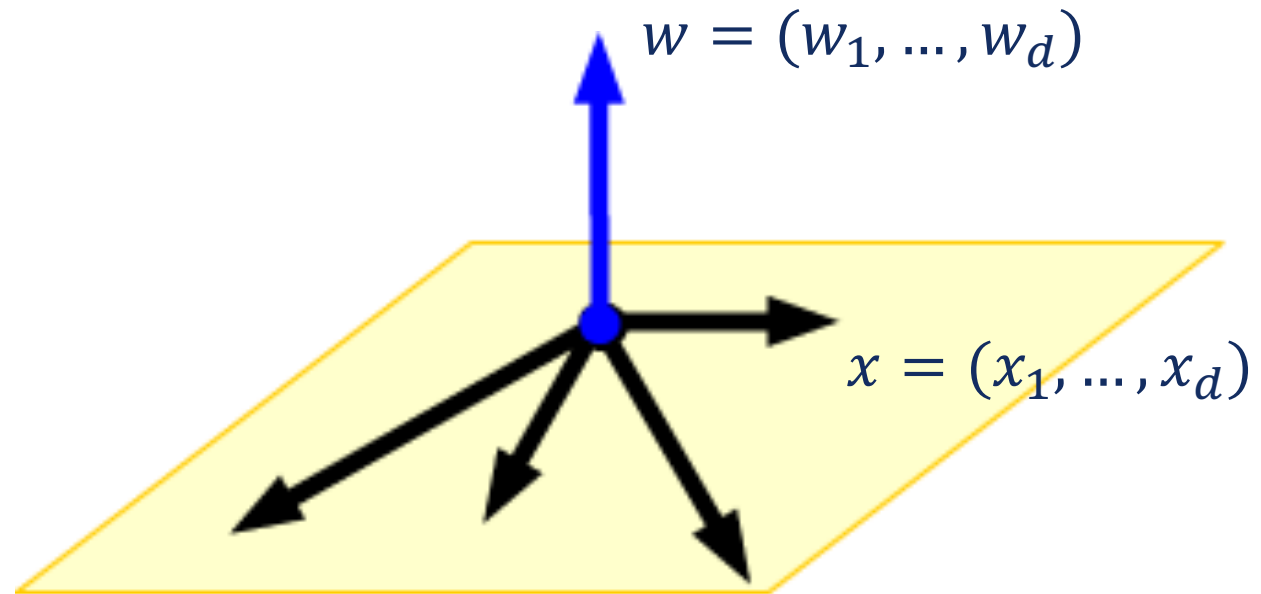
$$\alpha = 90^\circ$$

Линейные модели: формализация

Геометрическая интерпретация

$$a(x) = \langle w, x \rangle + w_0$$

Веса задают гиперплоскость, разделяющую классы.
Откуда берется гиперплоскость?

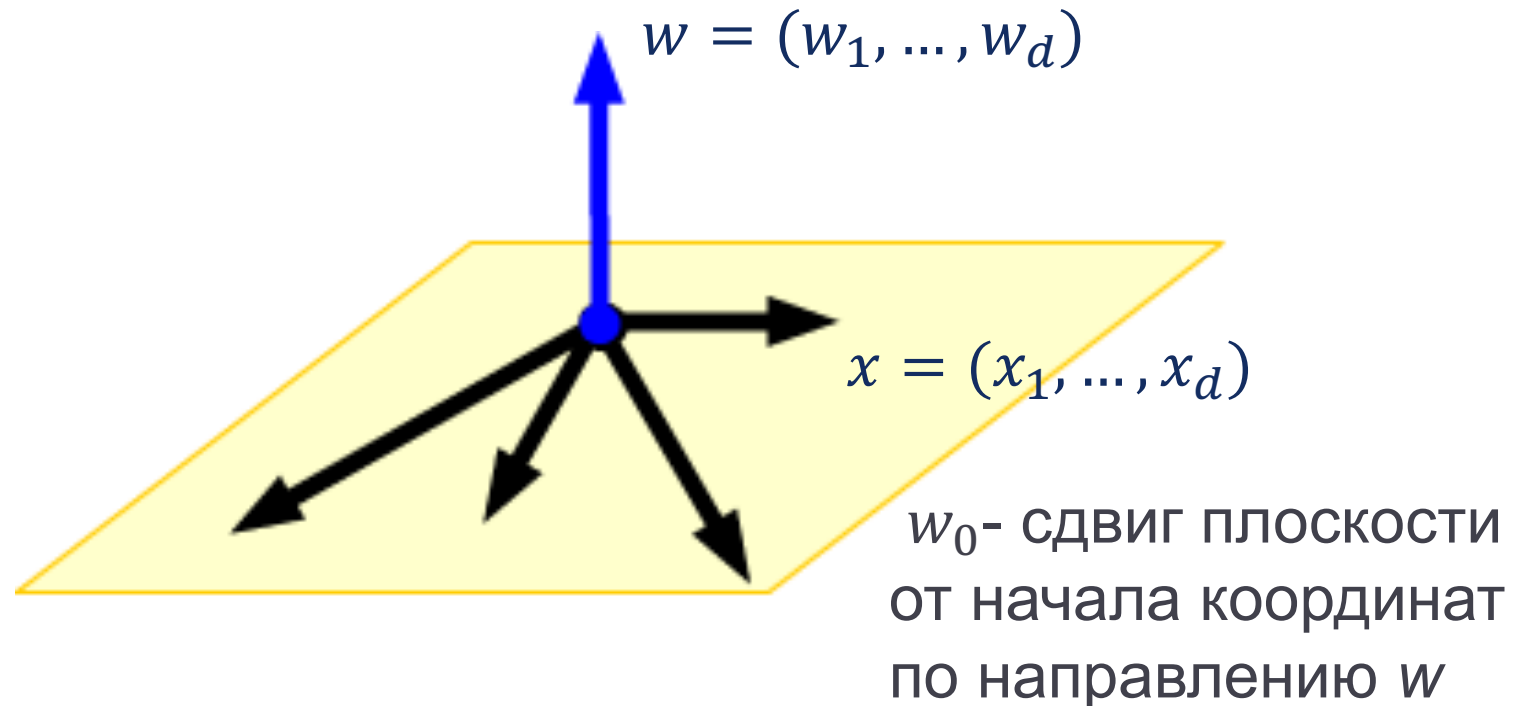


Линейные модели: формализация

Геометрическая интерпретация

$$a(x) = \langle w, x \rangle + w_0$$

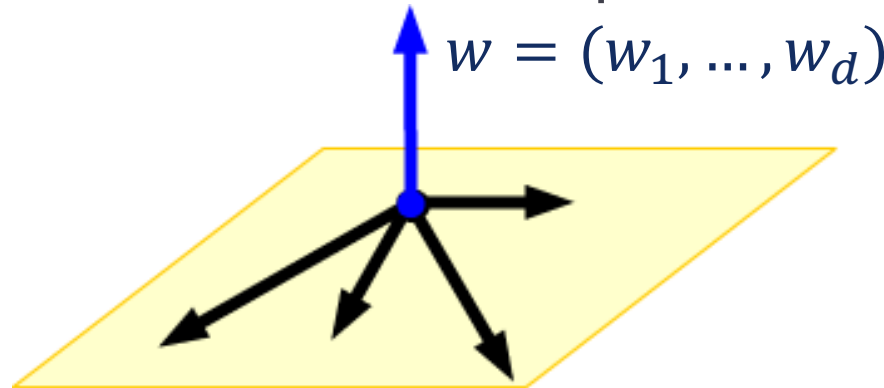
Веса задают гиперплоскость, разделяющую классы.
Откуда берется гиперплоскость?



Геометрическая интерпретация

$$a(x) = \langle w, x \rangle + w_0$$

Веса задают гиперплоскость, разделяющую классы.



- Если объект “над” гиперплоскостью, то его вектор и вектор w смотрят в одну сторону, скалярное произведение положительное
- Если объект с другой стороны от гиперплоскости — скалярное произведение отрицательное

Понятие отступа (margin)

Отступом алгоритма $a(x) = \text{sign}(f(x))$ на объекте x_i называется величина

$$M_i = y_i f(x_i)$$

(y_i - класс, к которому относится x_i)

$$\begin{aligned} M_i \leq 0 &\Leftrightarrow y_i \neq a(x_i) \\ M_i > 0 &\Leftrightarrow y_i = a(x_i) \end{aligned}$$

Понятие отступа (margin)

Отступом алгоритма $a(x) = \text{sign}(f(x))$ на объекте x_i называется величина

$$M_i = y_i f(x_i)$$

(y_i - класс, к которому относится x_i)

$$\begin{aligned} M_i \leq 0 &\Leftrightarrow y_i \neq a(x_i) \\ M_i > 0 &\Leftrightarrow y_i = a(x_i) \end{aligned}$$

Чем больше M_i — тем увереннее классификация

Линейные
модели:
формализация

Зная отступы, оценим потери

$$Q(a(w)) = \sum_{i=1}^n [M_i(w) < 0]$$

Линейные
модели:
формализация

Зная отступы, оценим потери

$$Q(a(w)) = \sum_{i=1}^n [M_i(w) < 0]$$

- $Q(a(w))$ – эмпирические риск
- хотелось бы найти такие w , чтобы $Q(a(w)) \rightarrow \min$

Зная отступы, оценим потери

$$Q(a(w)) = \sum_{i=1}^n [M_i(w) < 0] \leq \tilde{Q}(a(w)) = \sum_{i=1}^n L_i(M(w)) \rightarrow \min$$

Линейные
модели:
формализация

- $Q(a(w))$ – эмпирические риск
- $\tilde{Q}(a(w))$ – функция потерь

Линейные модели: формализация

Зная отступы, оценим потери

$$Q(a(w)) = \sum_{i=1}^n [M_i(w) < 0] \leq \tilde{Q}(a(w)) = \sum_{i=1}^n L_i(M(w)) \rightarrow \min$$

- $Q(a(w))$ – эмпирические риск
- $\tilde{Q}(a(w))$ - функция потерь

Теперь вместо задачи дискретной оптимизации мы решаем задачу непрерывной оптимизации! А значит можем использовать удобный аппарат для оптимизации (т.е. для обучения модели).

Линейные модели: построение

Линейные модели: построение

Задача оптимизации

$$Q(a(w)) = \sum_{i=1}^n L(M_i(w)) \rightarrow \min$$

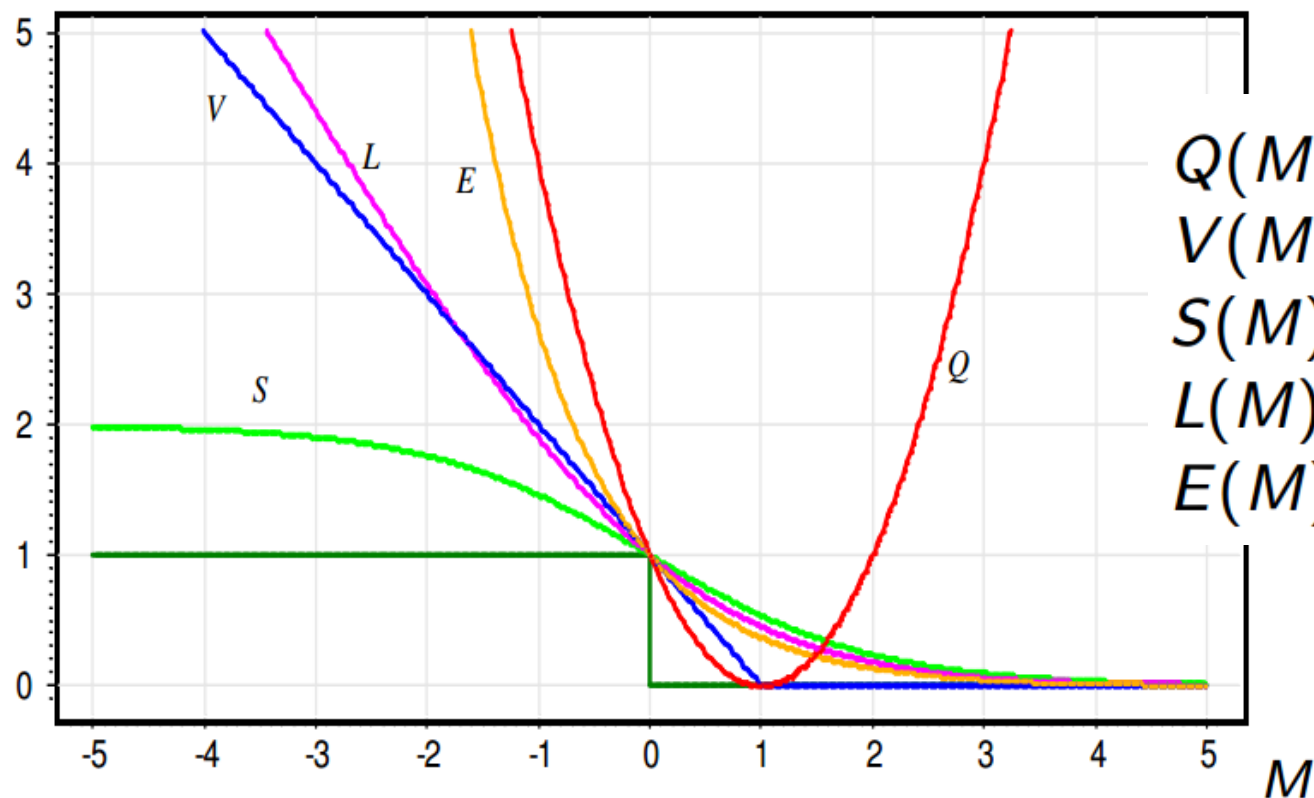
$$Q(a(w)) = \sum_{i=1}^n L(y_i, a_i) \rightarrow \min$$

Общий вид задачи оптимизации задан, остается несколько степеней свободы:

- функция потерь (L)
- метод оптимизации
- дополнительные ограничения

Функция потерь

Линейные
модели:
построение



$$\begin{aligned} Q(M) &= (1 - M)^2 \\ V(M) &= (1 - M)_+ \\ S(M) &= 2(1 + e^M)^{-1} \\ L(M) &= \log_2(1 + e^{-M}) \\ E(M) &= e^{-M} \end{aligned}$$

Линейные модели: построение

Задача оптимизации

$$Q(a(w)) = \sum_{i=1}^n L(M_i(w)) \rightarrow \min$$

$$Q(a(w)) = \sum_{i=1}^n L(y_i, a_i) \rightarrow \min$$

Общий вид задачи оптимизации задан, остается несколько степеней свободы:

- функция потерь (L)
- **метод оптимизации**
- дополнительные ограничения

Линейные
модели:
построение

Задача оптимизации

$$a(x) = \langle w, x \rangle + w_0$$

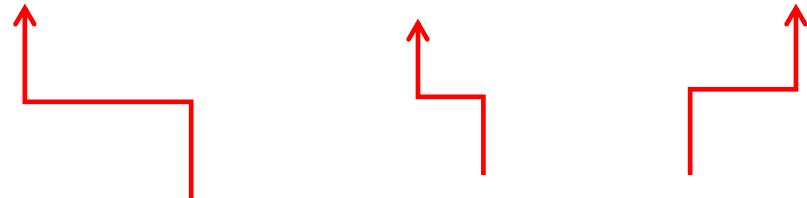
$$Q = \sum_{i=1}^l L(y_i, a(x_i)) \rightarrow \min_w$$

Линейные
модели:
построение

Решение задачи оптимизации

Матричная запись

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_l \end{pmatrix} \approx \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_l \end{pmatrix} = \begin{pmatrix} x_1^T \\ x_2^T \\ \dots \\ x_l^T \end{pmatrix} w$$


$$y \approx \hat{y} = Fw$$

$$w = \underset{w}{\operatorname{argmin}} \|y - \hat{y}\|^2$$

Линейные модели: построение

Решение задачи оптимизации

$$\frac{\partial (Fw - y)^2}{\partial w} = 2F^T (Fw - y) = 0$$

$$F^T Fw = F^T y$$

$$w = (F^T F)^{-1} F^T y$$

Нюансы:

- обращение матриц — вычислительно тяжелая операция
- более того, может не получиться: плохо обусловленная матрица, коррелирующие признаки

Линейные
модели:
построение

Задача оптимизации

$$a(x) = \langle w, x \rangle + w_0$$

$$Q = \sum_{i=1}^l L(y_i, a(x_i)) + \gamma V(w) \rightarrow \min_w$$

Гребневая регрессия
(Ridge regression):

$$V(w) = \|w\|_{l_2}^2 = \sum_{n=1}^d w_n^2$$

LASSO (least absolute
shrinkage and selection
operator):

$$V(w) = \|w\|_{l_1} = \sum_{n=1}^d |w_n|$$

Линейные
модели:
построение

Решение задачи оптимизации

Если добавить ℓ_2 регуляризацию

$$\frac{\partial (Fw - y)^2 + \gamma w^2}{\partial w} = 2F^T(Fw - y) + 2\gamma w = 0$$

$$(F^T F + \gamma I)w = F^T y$$

$$w = (F^T F + \gamma I)^{-1} F^T y$$

Решение задачи оптимизации

Если добавить ℓ_2 регуляризацию

$$\frac{\partial (Fw - y)^2 + \gamma w^2}{\partial w} = 2F^T(Fw - y) + 2\gamma w = 0$$

$$(F^T F + \gamma I)w = F^T y$$

$$w = (F^T F + \gamma I)^{-1} F^T y$$

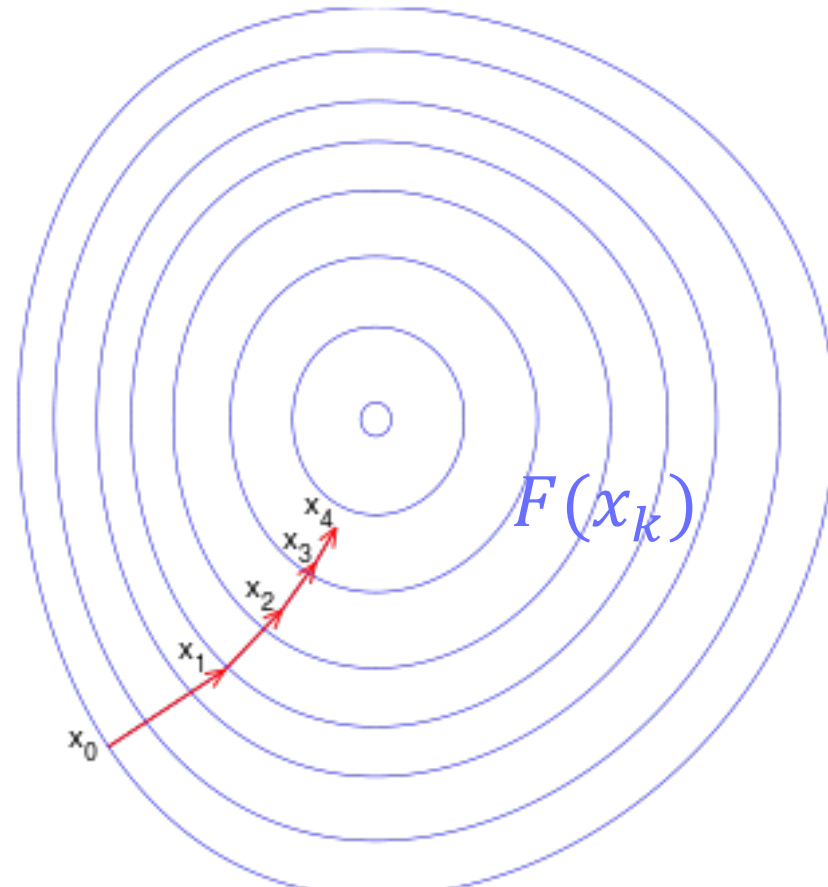
Но вычислительно задача всё равно сложная

Линейные модели: построение

Решение задачи оптимизации

Численное решение возможно с помощью
градиентного спуска (GD, Gradient Decent)

$$x_{k+1} = x_k - \gamma_k \nabla F(x_k)$$

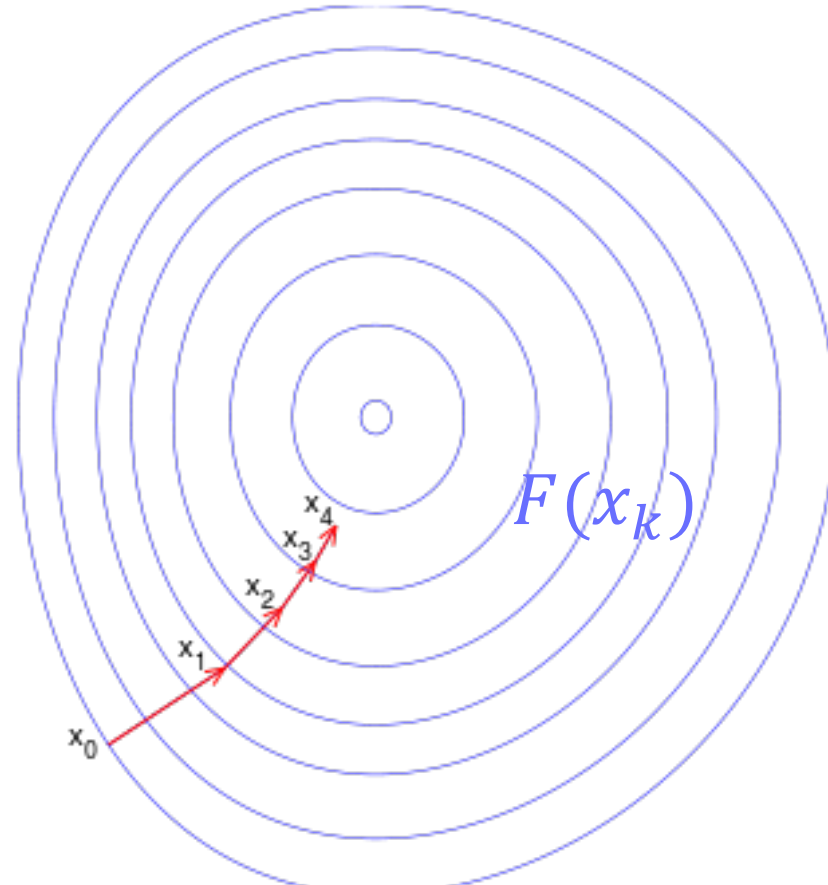


Линейные
модели:
построение

Градиентный спуск (GD)

$$\nabla_w \tilde{Q} = \sum_{i=1}^l \nabla L(M_i) = \sum_{i=1}^l L'(M_i) \frac{\partial M_i}{\partial w}$$

$$x_{k+1} = x_k - \gamma_k \nabla F(x_k)$$



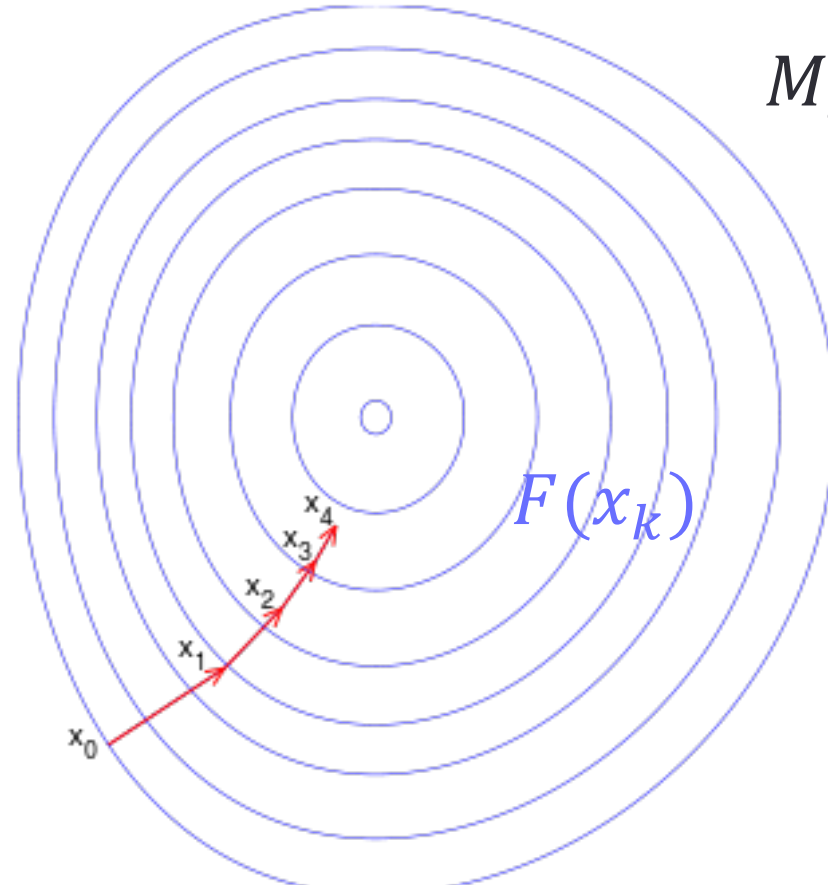
Линейные модели: построение

Градиентный спуск (GD)

$$\nabla_w \tilde{Q} = \sum_{i=1}^l \nabla L(M_i) = \sum_{i=1}^l L'(M_i) \frac{\partial M_i}{\partial w}$$

$$x_{k+1} = x_k - \gamma_k \nabla F(x_k)$$

$$M_i = y_i \langle w, x_i \rangle \Rightarrow \frac{\partial M_i}{\partial w} = y_i x_i$$



Линейные модели: построение

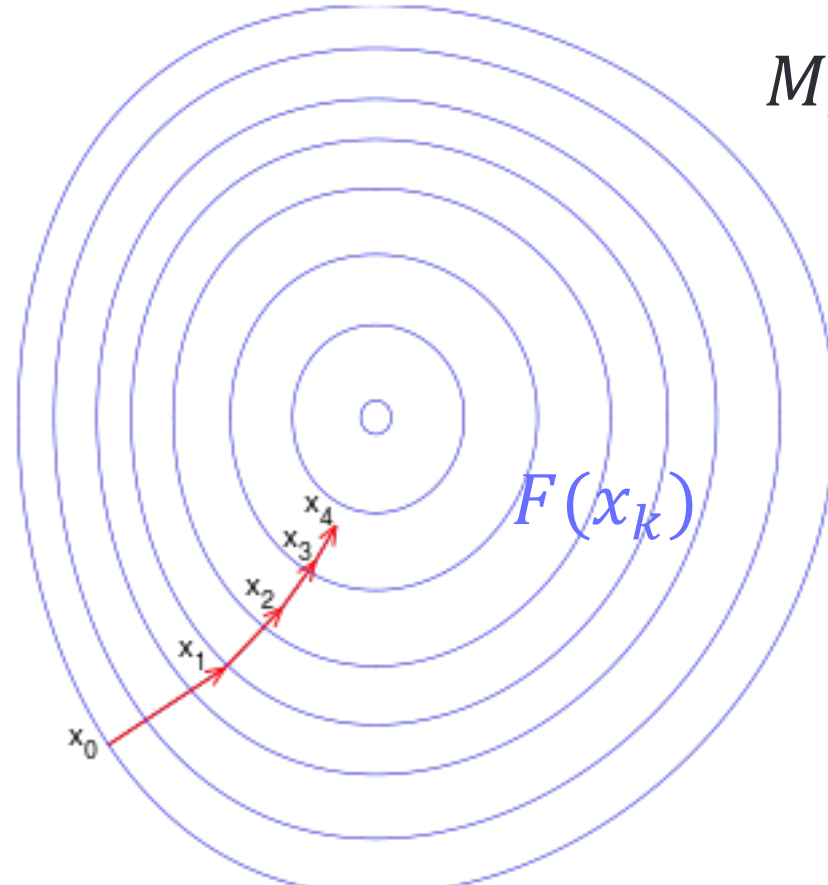
Градиентный спуск (GD)

$$\nabla_w \tilde{Q} = \sum_{i=1}^l \nabla L(M_i) = \sum_{i=1}^l L'(M_i) \frac{\partial M_i}{\partial w}$$

$$x_{k+1} = x_k - \gamma_k \nabla F(x_k)$$

$$M_i = y_i \langle w, x_i \rangle \Rightarrow \frac{\partial M_i}{\partial w} = y_i x_i$$

$$\nabla \tilde{Q} = \sum_{i=1}^l y_i x_i L'(M_i)$$



Линейные модели: построение

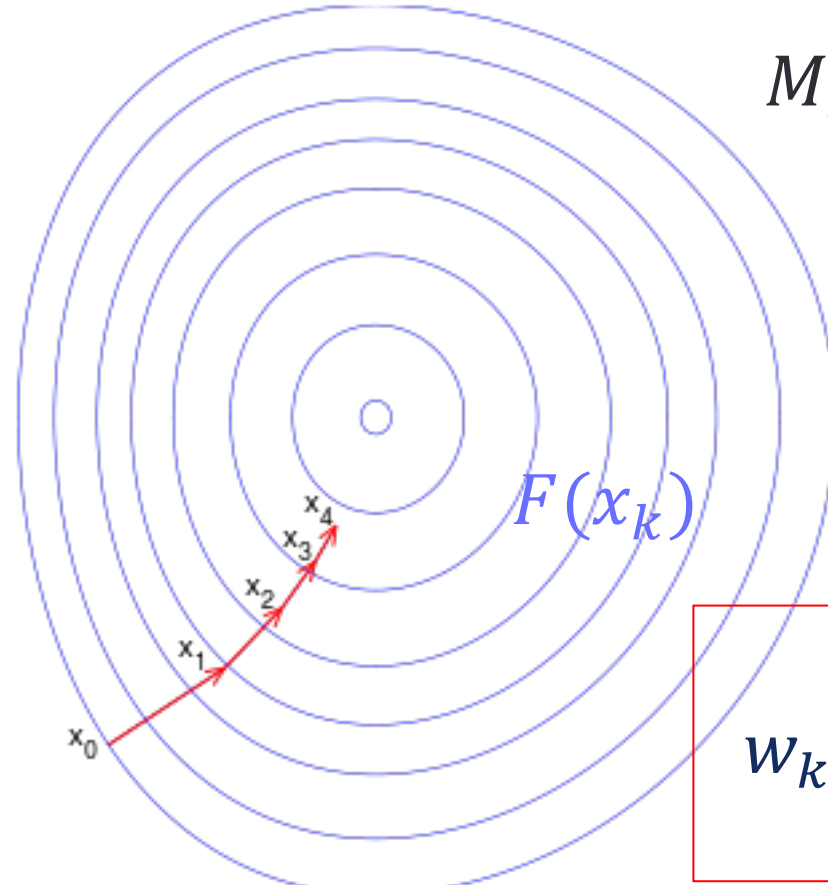
Градиентный спуск (GD)

$$\nabla_w \tilde{Q} = \sum_{i=1}^l \nabla L(M_i) = \sum_{i=1}^l L'(M_i) \frac{\partial M_i}{\partial w}$$

$$x_{k+1} = x_k - \gamma_k \nabla F(x_k)$$

$$M_i = y_i \langle w, x_i \rangle \Rightarrow \frac{\partial M_i}{\partial w} = y_i x_i$$

$$\nabla \tilde{Q} = \sum_{i=1}^l y_i x_i L'(M_i)$$



$$w_{k+1} = w_k - \gamma_k \sum_{i=1}^l y_i x_i L'(M_i)$$

Линейные
модели:
построение

Стохастический градиент (SGD)

$$w_{k+1} = w_k - \gamma_k \sum_{i=1}^l y_i x_i L'(M_i)$$

$$w_{k+1} = w_k - \gamma_k y_i x_i L'(M_i)$$

x_i — случайный элемент обучающей выборки

Линейные модели: построение

Стохастический градиент (SGD)

$$w_{k+1} = w_k - \gamma_k \sum_{i=1}^l y_i x_i L'(M_i)$$

$$w_{k+1} = w_k - \gamma_k y_i x_i L'(M_i)$$

x_i — случайный элемент обучающей выборки

Почему это вообще работает?

- да, траектория спуска будет со скачками
- шаги не такие качественные как в GD, но очень быстрые! Значит можно быстро сделать много шагов

Линейные модели: построение

Стохастический градиент (SGD)

$$w_{k+1} = w_k - \gamma_k \sum_{i=1}^l y_i x_i L'(M_i)$$

$$w_{k+1} = w_k - \gamma_k y_i x_i L'(M_i)$$

x_i — случайный элемент обучающей выборки

Ускорение оптимизации:

- выбор начального приближения
- адаптивный шаг
- порядок предъявления объектов
- расчет градиента (например, с инерцией и пр.)

Линейные модели: построение

Задача оптимизации

$$Q(a(w)) = \sum_{i=1}^n L(M_i(w)) \rightarrow \min$$

$$Q(a(w)) = \sum_{i=1}^n L(y_i, a_i) \rightarrow \min$$

Общий вид задачи оптимизации задан, остается несколько степеней свободы:

- функция потерь (L)
- метод оптимизации
- дополнительные ограничения

Задача оптимизации

Линейные
модели:
построение

$$a(x) = \langle w, x \rangle + w_0$$

$$Q = \sum_{i=1}^l L(y_i, a(x_i)) \rightarrow \min_w$$

Задача оптимизации

Линейные
модели:
построение

$$a(x) = \langle w, x \rangle + w_0$$

$$Q = \sum_{i=1}^l L(y_i, a(x_i)) \rightarrow \min_w$$

$$V(w) = \|w\|_{l_2}^2 = \sum_{n=1}^d w_n^2$$

$\ell 2$ -регуляризация

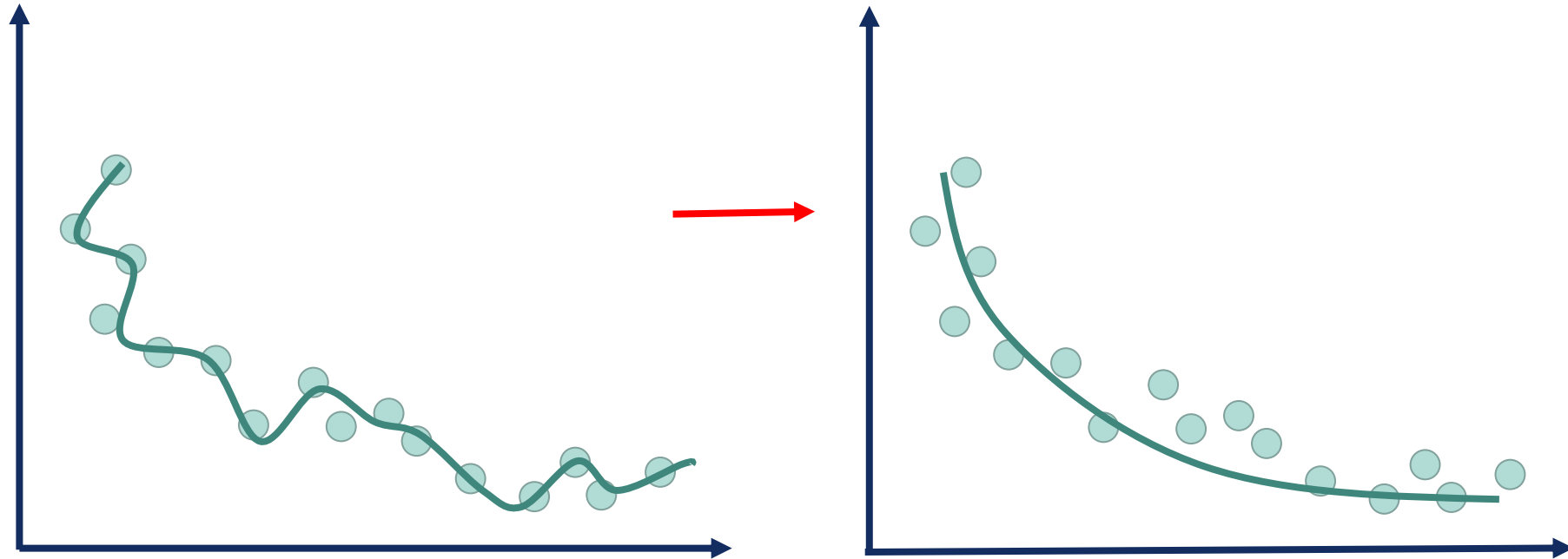
$$V(w) = \|w\|_{l_1} = \sum_{n=1}^d |w_n|$$

$\ell 1$ -регуляризация

Линейные модели: построение

Регуляризация

Регуляризация – способ наложить дополнительные ограничения на веса линейной модели



Линейные модели: построение

Регуляризация

$$\left\{ \begin{array}{l} \tilde{Q} = \sum_{i=1}^l L(M_i) \rightarrow \min \\ \sum_{n=1}^d |w_n| \leq \tau \end{array} \right.$$

$\ell 1$ -регуляризация

$$\left\{ \begin{array}{l} \tilde{Q} = \sum_{i=1}^l L(M_i) \rightarrow \min \\ \sum_{n=1}^d w_n^2 \leq \tau \end{array} \right.$$

$\ell 2$ -регуляризация

Линейные
модели:
построение

Регуляризация

$$\sum_{i=1}^l L(M_i) + \gamma \sum_{n=1}^d |w_n| \rightarrow \min$$

$$\sum_{i=1}^l L(M_i) + \gamma \sum_{n=1}^d w_n^2 \rightarrow \min$$

Линейные модели: построение

Регуляризация

$$\sum_{i=1}^l L(M_i) + \gamma \sum_{n=1}^d |w_n| \rightarrow \min$$

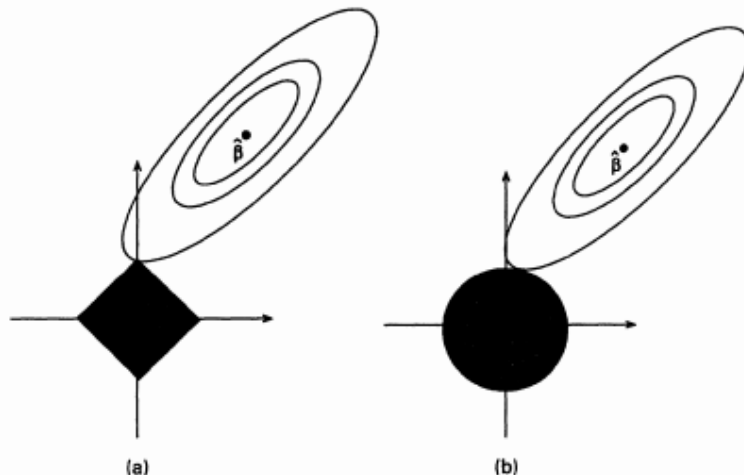
$$\sum_{i=1}^l L(M_i) + \gamma \sum_{n=1}^d w_n^2 \rightarrow \min$$

- Разреженность – ℓ_1 -регуляризация делает вектор весов более **разреженным** (содержащим больше нулей)
- В случае линейной классификации это означает **отбор признаков**: признаки с нулевыми весами не используются в классификации
- ℓ_2 -регуляризация приводит к получению вектора с весами, меньшими по модулю

Линейные модели: построение

Регуляризация

- Разреженность — ℓ_1 -регуляризация приводит к получению более **разреженного** вектора весов (содержащим больше нулей)
- В случае линейной классификации это означает **отбор признаков**: признаки с нулевыми весами не используются в классификации
- ℓ_2 -регуляризация приводит к получению вектора с весами, меньшими по модулю



Часто используемые
линейные модели

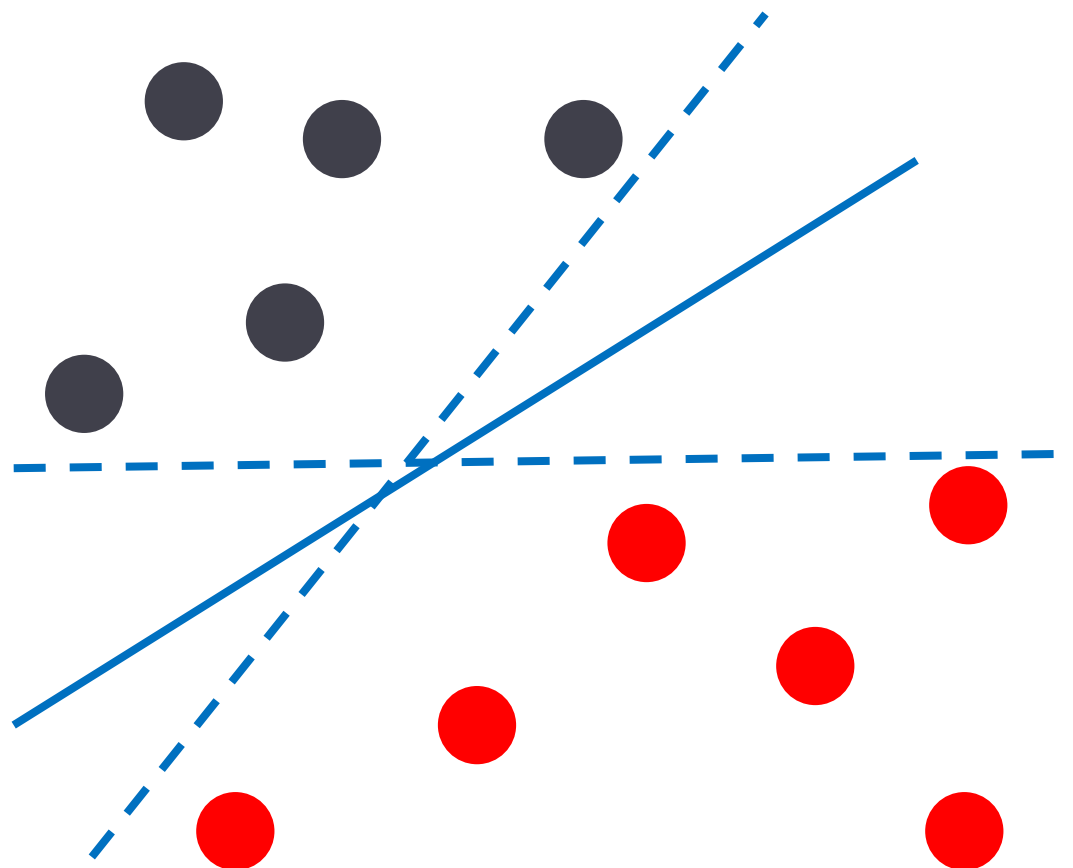
Конструирование линейных моделей

Линейные модели: построение

Классификатор	Функция потерь	Регуляризатор
SVM (Support vector machine, метод опорных векторов)	$L(M) = \max\{0, 1 - M\} = (1 - M)_+$	$\sum_{k=1}^m w_k^2$
Логистическая регрессия	$L(M) = \log(1 + e^{-M})$	$\sum_{k=1}^m w_k^2 / \sum_{k=1}^m w_k $

Линейные
модели: SVM

SVM: линейно разделимые
выборки



Линейные модели: SVM

SVM: линейно разделимые выборки

$$a(x) = \langle w, x \rangle + w_0$$

При умножении весов на $c > 0$ модель не изменится
(знак скалярных произведений сохранится)

Нормируем веса модели:

$$\min_{x \in X} |\langle w, x \rangle + w_0| = 1$$

Строим $a(x)$ так, чтобы расстояния от разделяющей плоскости до объектов было как можно больше:

$$\rho(x, a) = \frac{|\langle w, x \rangle + w_0|}{\|w\|}$$

Линейные модели: SVM

SVM: линейно разделимые выборки

Чему равно расстояние от плоскости до ближайшего объекта?

$$\min_{x \in X} \rho(x, a) = \min_{x \in X} \frac{|\langle w, x \rangle + w_0|}{\|w\|} = \frac{1}{\|w\|} * 1$$

Выпишем задачу оптимизации, замети что $\min \|w\|$ можно заменить на $\min \|w\|^2$:

$$\begin{cases} \|w\|^2 \rightarrow \min_w \\ y_i(\langle w, x \rangle + w_0) \geq 1 \end{cases}$$

Неравенство обеспечивает минимальное расстояние до объектов > 1 и верную классификацию

Линейные модели: SVM

SVM: неразделимые выборки

Адаптируем для неразделимой выборки:

$$\begin{cases} ||w||^2 + c \sum_{i=1}^l \xi_i \rightarrow \min_{w, \xi} \\ y_i(< w, x > + w_0) \geq 1 - \xi_i \\ \xi_i > 0 \end{cases}$$

таким образом мы разрешаем делать ошибки, но не большие

Линейные модели: SVM

SVM: неразделимые выборки

Выпишем безусловную задачу:

мы минимизируем ξ_i и для них есть 2 ограничения снизу. Возьмём максимальное из них

$$||w||^2 + c \sum_{i=1}^l \max(0, 1 - y_i(< w, x_i >, w_0)) \rightarrow \min_w$$

Сравним с задачей для SVM:

$$\sum_{i=1}^l \max(0, 1 - M_i) + c||w||^2 \rightarrow \min_w$$

Линейные модели: лог. регрессия

Логистическая регрессия

Предположим, что данные имеют вероятностную природу

$$p(y = 1 | x)$$

А модель выдает оценку вероятности $a(x) = p$

Запишем правдоподобие выборки

$$\prod_{i=1}^l a(x_i)^{[y_i=1]} (1 - a(x_i))^{[y_i=-1]} \rightarrow \max$$

Прологарифмируем

$$-\sum_{i=1}^l ([y = 1] \log a(x_i) + [y = -1] \log(1 - a(x_i))) \rightarrow \min$$

Линейные модели: лог. регрессия

Логистическая регрессия

$$-\sum_{i=1}^l ([y = 1] \log a(x_i) + [y = -1] \log(1 - a(x_i))) \rightarrow \min$$

Видно, что функция потерь

$$L(y, a) = -[y = 1] \log a(x) - [y = -1] \log(1 - a(x))$$

Выберем $a(x)$ с областью значений $[0;1]$

$$a(w, x) = \sigma(< w, x >) = \frac{1}{1 + e^{-< w, x >}}$$

Остается подставить $a(x)$ в функцию потерь $L(y, a)$ и оценить потери модели $Q(a)$

Линейные модели: лог. регрессия

Логистическая регрессия

$$-\sum_{i=1}^l \left([y = 1] \log \frac{1}{1 + e^{-\langle w, x \rangle}} + [y = -1] \log \left(1 - \frac{1}{1 + e^{-\langle w, x \rangle}} \right) \right) \rightarrow \min$$

преобразуем второй логарифм

$$-\sum_{i=1}^l \left([y = 1] \log \frac{1}{1 + e^{-\langle w, x \rangle}} + [y = -1] \log \left(\frac{1}{1 + e^{\langle w, x \rangle}} \right) \right) \rightarrow \min$$

мало

перепишем более компактно

$$-\sum_{i=1}^l \log \frac{1}{1 + e^{-y_i \langle w, x_i \rangle}} \rightarrow \min$$

это и есть логистическая регрессия, полученная из вероятностных соображений

$$\sum_{i=1}^l \log(1 + e^{-y_i \langle w, x_i \rangle}) \rightarrow \min$$

Многоклассовые модели

В случае нескольких классов задача сводится к обучению n классификаторов one-vs-all и выбору \max

$$a(x) = \arg \max_{y \in Y} (< w_y, x >)$$

Для вероятностной классификации применяется функция SoftMax для расчета вероятностей

$$p(y|x, w) = \frac{e^{<w_y, x>}}{\sum_z^Y e^{<w_z, x>}} = \text{SoftMax } < w_y, x >$$

Линейные модели: построение

To take away:

Линейные модели имеют ряд преимуществ:

- Простота реализации
- Скорость работы
- Легкое обновление, переобучение в production
- Обучение на больших данных, множество адаптаций
- Хорошее качество, когда много признаков
- Приемлемое качество, когда данных мало

Особенности применения:

- Модель может оказаться слишком простой для задачи
- Требуется бороться с переобучением:
регуляризация, масштабирование признаков

Машинное обучение: линейные модели классификации и регрессии

Спасибо!
Эмили Драль