Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Д. П. Ветров[1] Д. А. Кропотов[2]

[1]МГУ, ВМиК, каф. ММП

[2]ВЦ РАН

Спецкурс «Байесовские методы машинного обучения»

План лекции

Ликбез

- Нормальное распределение
- Решение нерешаемых СЛАУ

Статистическая постановка задачи машинного обучения

- Вероятностное описание
- Байесовские решающие правила

Линейная регрессия

- Классическая линейная регрессия
- Метод наименьших квадратов
- Вероятностная постановка задачи

Задача классификации

- Логистическая регрессия
- Метод IRLS

План лекции

Ликбез
Нормальное распределение
Решение нерешаемых СЛАУ
Статистическая постановка задачи машинного обучения
Вероятностное описание
Байесовские решающие правила
Линейная регрессия
Классическая линейная регрессия
Метод наименьших квадратов
Вероятностная постановка задачи
Задача классификации
Логистическая регрессия
Метод IRLS

Нормальное распределение

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Нормальное распределение

Решение нерешаемых СЛАУ

Статистическая постановка задачи машинного обучения

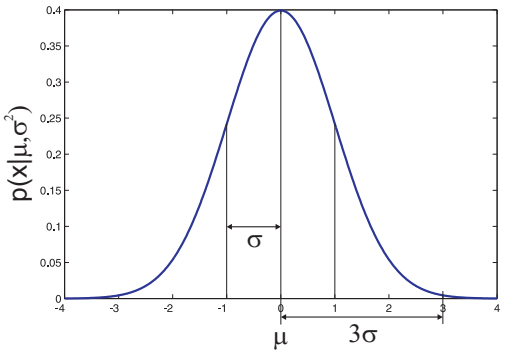
Линейная регрессия

Задача классификации

- Нормальное распределение играет важнейшую роль в математической статистике

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

$$\mu = \mathbb{E}X, \quad \sigma^2 = \mathbb{D}X \triangleq \mathbb{E}(X-\mathbb{E}X)^2$$



- Из центральной предельной теоремы следует, что сумма независимых случайных величин с ограниченной дисперсией стремится к нормальному распределению
- На практике, многие случайные величины можно считать приближенно нормальными

Многомерное нормальное распределение

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Нормальное распределение

Решение нерешаемых СЛАУ

Статистическая постановка задачи машинного обучения

Линейная регрессия

Задача классификации

- Многомерное нормальное распределение имеет вид

$$X \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi}^n \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right),$$

  где $\boldsymbol{\mu} = \mathbb{E}X$, $\Sigma = \mathbb{E}(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^T$ — вектор математических ожиданий каждой из n компонент и матрица ковариаций соответственно
- Матрица ковариаций показывает, насколько сильно связаны (коррелируют) компоненты многомерного нормального распределения

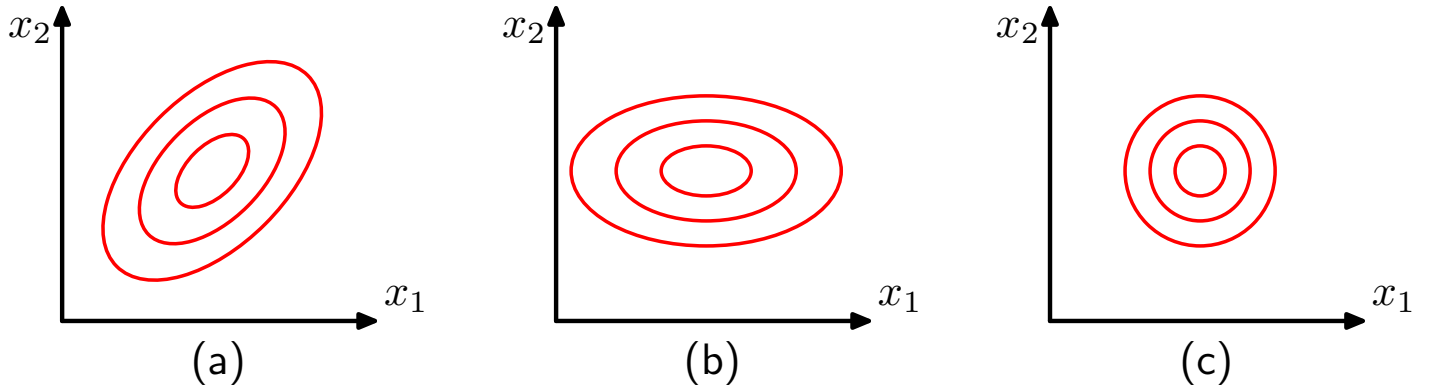
$$\Sigma_{ij} = \mathbb{E}(X_i - \mu_i)(X_j - \mu_j) = \mathrm{Cov}(X_i, X_j)$$

- Если мы поделим ковариацию на корень из произведений дисперсий, то получим коэффициент корреляции

$$\rho(X_i, X_j) \triangleq \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathbb{D}X_i \mathbb{D}X_j}} \in [-1, 1]$$

Особенности нормального распределения

- Нормальное распределение полностью задается первыми двумя моментами (мат. ожидание и матрица ковариаций/дисперсия)
- Матрица ковариаций неотрицательно определена, причем на диагоналях стоят дисперсии соответствующих компонент
- Нормальное распределение имеет очень легкие хвосты: большие отклонения от мат. ожидания практически невозможны. Это обстоятельство нужно учитывать при приближении произвольных случайных величин нормальными



План лекции

Ликбез
Нормальное распределение
Решение нерешаемых СЛАУ
Статистическая постановка задачи машинного обучения
Вероятностное описание
Байесовские решающие правила
Линейная регрессия
Классическая линейная регрессия
Метод наименьших квадратов
Вероятностная постановка задачи
Задача классификации
Логистическая регрессия
Метод IRLS

Псевдообращение матриц

- Предположим, нам необходимо решить СЛАУ вида $A\boldsymbol{x} = \boldsymbol{b}$
- Если бы матрица A была квадратной и невырожденной (число уравнений равно числу неизвестных и все уравнения линейно независимы), то решение задавалось бы формулой $\boldsymbol{x} = A^{-1}\boldsymbol{b}$
- Предположим, что число уравнений больше числа неизвестных, т.е. матрица A прямоугольная. Домножим обе части уравнения на A^T слева

$$A^T A\boldsymbol{x} = A^T\boldsymbol{b}$$

- В левой части теперь квадратная матрица и ее можно перенести в правую часть

$$\boldsymbol{x} = \left(A^T A\right)^{-1} A^T\boldsymbol{b}$$

- Операция $\left(A^T A\right)^{-1} A^T$ называется псевдообращением матрицы A, а $\boldsymbol{x}$ – псевдорешением

Нормальное псевдорешение

- Если матрица A^TA вырождена, псевдорешений бесконечно много, причем найти их на компьютере нетривиально
- Для решения этой проблемы используется ридж-регуляризация матрицы A^TA

$$A^TA + \lambda I,$$

где I – единичная матрица, а λ – коэффициент регуляризации. Такая матрица невырождена для любых $\lambda > 0$

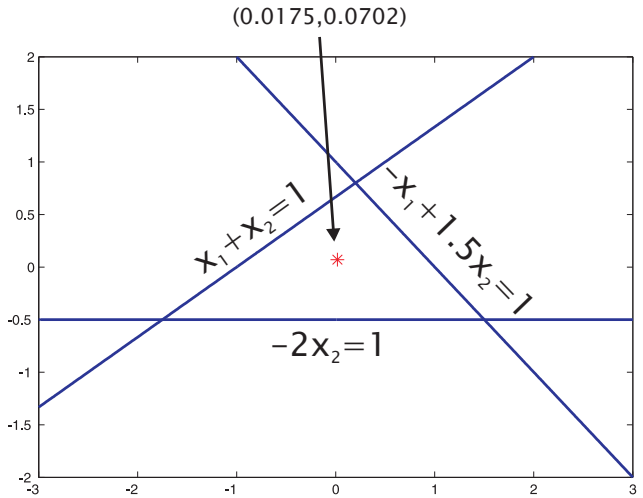
- Величина

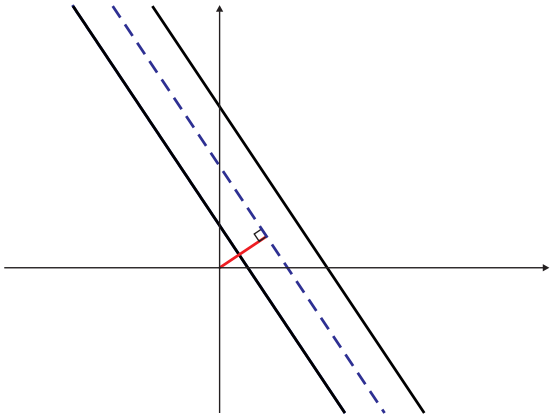
$$\boldsymbol{x} = \left(A^TA + \lambda I\right)^{-1} A^T\boldsymbol{b}$$

называется нормальным псевдорешением. Оно всегда единственно и при небольших положительных λ определяет псевдорешение с наименьшей нормой

Графическая иллюстрация

- Псевдорешение соответствует точке, минимизирующей невязку, а нормальное псевдорешение отвечает псевдорешению с наименьшей нормой





- Заметим, что псевдообратная матрица $\left(A^TA\right)^{-1}A^T$ совпадает с обратной матрицей A^{-1} в случае невырожденных квадратных матриц

План лекции

Ликбез

Статистическая постановка задачи машинного обучения

Вероятностное описание

Байесовские решающие правила

Линейная регрессия

Задача классификации

Ликбез
- Нормальное распределение
- Решение нерешаемых СЛАУ

Статистическая постановка задачи машинного обучения
- Вероятностное описание
- Байесовские решающие правила

Линейная регрессия
- Классическая линейная регрессия
- Метод наименьших квадратов
- Вероятностная постановка задачи

Задача классификации
- Логистическая регрессия
- Метод IRLS

Основные обозначения

- В дальнейшем будут рассматриваться преимущественно задачи классификации и восстановления регрессии
- В этих задачах обучающая выборка представляет собой набор отдельных объектов $X = \{\boldsymbol{x}_i\}_{i=1}^n$, характеризующихся вектором вещественнозначных признаков $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})$
- Каждый объект также обладает скрытой переменной $t \in \mathcal{T}$
- Предполагается, что существует зависимость между признаками объекта и значением скрытой переменной
- Для объектов обучающей выборки значение скрытой переменной известно $\boldsymbol{t} = \{t_i\}_{i=1}^n$

Статистическая постановка задачи

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Статистическая постановка задачи машинного обучения

Вероятностное описание

Байесовские решающие правила

Линейная регрессия

Задача классификации

- Каждый объект описывается парой $(\boldsymbol{x}, t)$
- При статистической (вероятностной) постановке задачи машинного обучения предполагается, что обучающая выборка является набором независимых, одинаково распределенных случайных величин, взятых из некоторой генеральной совокупности
- В этом случае уместно говорить о плотности распределения объектов $p(\boldsymbol{x}, t)$ и использовать вероятностные термины (математическое ожидание, дисперсия, правдоподобие) для описания и решения задачи
- Заметим, что это не единственная возможная постановка задачи машинного обучения

Качество обучения

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Статистическая постановка задачи машинного обучения

Вероятностное описание

Байесовские решающие правила

Линейная регрессия

Задача классификации

- Качество обучения определяется точностью прогноза на генеральной совокупности
- Пусть $S(t, \hat{t})$ – функция потерь, определяющая штраф за прогноз $\hat{t}$ при истинном значении скрытой переменной t
- Разумно ожидать, что минимум этой функции достигается при $\hat{t} = t$
- Примерами могут служить $S_r(t, \hat{t}) = (t - \hat{t})^2$ для задачи восстановления регрессии и $S_c(t, \hat{t}) = I\{\hat{t} \neq t\}$ для задачи классификации

Абсолютный критерий качества

- Если бы функция $p(\boldsymbol{x}, t)$ была известна, задачи машинного обучения не существовало
- В самом деле абсолютным критерием качества обучения является мат. ожидание функции потерь, взятое по генеральной совокупности

$$\mathbb{E}S(t, \hat{t}) = \int S(t, \hat{t}(\boldsymbol{x}))p(\boldsymbol{x}, t)d\boldsymbol{x}dt \to \min,$$

  где $\hat{t}(\boldsymbol{x})$ – решающее правило, возвращающее величину прогноза для вектора признаков $\boldsymbol{x}$
- Вместо методов машинного обучения сейчас бы активно развивались методы оптимизации и взятия интегралов от функции потерь :)
- К сожалению (а может, к счастью), распределение объектов генеральной совокупности неизвестно, поэтому абсолютный критерий качества обучения не может быть подсчитан

План лекции

Ликбез

Статистическая постановка задачи машинного обучения

Вероятностное описание

Байесовские решающие правила

Линейная регрессия

Задача классификации

Ликбез
- Нормальное распределение
- Решение нерешаемых СЛАУ

Статистическая постановка задачи машинного обучения
- Вероятностное описание
- Байесовские решающие правила

Линейная регрессия
- Классическая линейная регрессия
- Метод наименьших квадратов
- Вероятностная постановка задачи

Задача классификации
- Логистическая регрессия
- Метод IRLS

Идеальный классификатор

- Итак, одна из основных задач теории машинного обучения — это разработка способов косвенного оценивания качества решающего правила и выработка новых критериев для оптимизации в ходе обучения
- Рассмотрим задачу классификации с функцией потерь вида $S_c(t,\hat{t}) = I\{\hat{t} \neq t\}$ и гипотетический классификатор $t_B(\boldsymbol{x}) = \arg\max_{t\in\mathcal{T}} p(\boldsymbol{x},t) = \arg\max_{t\in\mathcal{T}} p(t|\boldsymbol{x})$
- Справделива следующая цепочка неравенств

$$\mathbb{E}S(t,\hat{t}) = \int\int S(t,\hat{t}(\boldsymbol{x}))p(\boldsymbol{x},t)d\boldsymbol{x}dt =$$

$$\sum_{s=1}^{l}\int S(s,\hat{t}(\boldsymbol{x}))p(\boldsymbol{x},s)d\boldsymbol{x} = 1 - \int p(\boldsymbol{x},\hat{t}(\boldsymbol{x}))d\boldsymbol{x} \geq$$

$$\geq 1 - \int \max_t p(\boldsymbol{x},t)d\boldsymbol{x} = 1 - \int p(\boldsymbol{x},t_B(\boldsymbol{x}))d\boldsymbol{x} = \mathbb{E}S(t,t_B)$$

Идеальная регрессия

- Рассмотрим задачу восстановления регрессии с квадратичной функцией потерь вида $S_r(t,\hat{t}) = (t-\hat{t})^2$ и гипотетическое решающее правило

$$t_B(\boldsymbol{x}) = \mathbb{E}_{t|\boldsymbol{x}} t = \int t p(t|\boldsymbol{x}) dt$$

- Справделива следующая цепочка неравенств

$$\mathbb{E}S(t,\hat{t}) = \int\int S(t,\hat{t}(\boldsymbol{x}))p(\boldsymbol{x},t)d\boldsymbol{x}dt =$$

$$\int\int (t-\hat{t}(\boldsymbol{x}))^2 p(\boldsymbol{x},t)d\boldsymbol{x}dt = \int\int ((t-\mathbb{E}t)+(\mathbb{E}t-\hat{t}(\boldsymbol{x})))^2 p(\boldsymbol{x},t)d\boldsymbol{x}dt =$$

$$= \int\int (t-\mathbb{E}t)^2 p(\boldsymbol{x},t)d\boldsymbol{x}dt + 2\int\int (t-\mathbb{E}t)(\mathbb{E}t-\hat{t}(\boldsymbol{x}))p(\boldsymbol{x},t)d\boldsymbol{x}dt +$$

$$+ \int\int (\mathbb{E}t-\hat{t}(\boldsymbol{x}))^2 p(\boldsymbol{x},t)d\boldsymbol{x}dt \geq$$

$$\geq \int\int (t-\mathbb{E}t)^2 p(t|\boldsymbol{x})p(\boldsymbol{x})dtd\boldsymbol{x} = \mathbb{E}S(t,t_B(\boldsymbol{x}))$$

Особенности байесовских решающих правил

- Таким образом, знание распределения объектов генеральной совокупности приводит к получению оптимальных решающих правил в явной форме
- Такой оптимальные решающие правила называются байесовскими
- Если бы удалось с высокой точностью оценить значение условной плотности $p(t|\boldsymbol{x})$ для всех $\boldsymbol{x}$ и t, обе основные задачи машинного обучения можно было считать решенными
- На этом основан один из существующих подходов к машинному обучению

План лекции

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

Классическая линейная регрессия

Метод наименьших квадратов

Вероятностная постановка задачи

Задача классификации

Ликбез
- Нормальное распределение
- Решение нерешаемых СЛАУ

Статистическая постановка задачи машинного обучения
- Вероятностное описание
- Байесовские решающие правила

Линейная регрессия
- Классическая линейная регрессия
- Метод наименьших квадратов
- Вероятностная постановка задачи

Задача классификации
- Логистическая регрессия
- Метод IRLS

Задача восстановления регрессии

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

Классическая линейная регрессия

Метод наименьших квадратов

Вероятностная постановка задачи

Задача классификации

- Задача восстановления регрессии предполагает наличие связи между наблюдаемыми признаками $\boldsymbol{x}$ и непрерывной переменной t
- В отличие от задачи интерполяции допускаются отклонения решающего правила от правильных ответов на объектах обучающей выборки
- Уравнение регрессии $y(\boldsymbol{x}, \boldsymbol{w})$ ищется в некотором параметрическом виде путем нахождения наилучшего значения вектора весов

$$\boldsymbol{w}_* = \arg\max_{\boldsymbol{w}} F(X, \boldsymbol{t}, \boldsymbol{w})$$

Линейная регрессия

- Наиболее простой и изученной является линейная регрессия
- Главная особенность: настраиваемые параметры входят в решающее правило линейно
- Заметим, что линейная регрессия не обязана быть линейной по признакам
- Общее уравнение регрессии имеет вид

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

Особенность выбора базисных функций

- Общего метода выбора базисных функций $\phi_j(\boldsymbol{x})$ — не существует
- Обычно они подбираются из априорных соображений (например, если мы пытаемся восстановить какой-то периодический сигнал, разумно взять функции тригонометрического ряда) или путем использования некоторых «универсальных» базисных функций
- Наиболее распространенными базисными функциями являются
  - $\phi(\boldsymbol{x}) = x_k$
  - $\phi(\boldsymbol{x}) = x_{k_1} x_{k_2} \ldots x_{k_l}$
  - $\phi(\boldsymbol{x}) = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{x}_0\|^p), \ \gamma, p > 0.$
- Метод построения линейной регрессии (настройки весов $\boldsymbol{w}$) не зависит от выбора базисных функций

Формализация задачи

- Пусть $S(t, \hat{t})$ — функция потерь от ошибки в определении регрессионной переменной t
- Необходимо минимизировать потери от ошибок на генеральной совокупности

$$\mathbb{E}S(t, y(\boldsymbol{x}, \boldsymbol{w})) = \int\int S(t, y(\boldsymbol{x}, \boldsymbol{w}))p(\boldsymbol{x}, t)d\boldsymbol{x}dt \to \min_{\boldsymbol{w}}$$

- Дальнейшие рассуждения зависят от вида функции потерь
- Во многих случаях даже не нужно восстанавливать полностью условное распределение $p(t|\boldsymbol{x})$

Важная теорема

- Теорема. Пусть функция потерь имеет вид
  - $S(t, \hat{t}) = (t - \hat{t})^2$ — «Потери старушки»;
  - $S(t, \hat{t}) = |t - \hat{t}|$ — «Потери олигарха»;
  - $S(t, \hat{t}) = \delta^{-1}(t - \hat{t})$ — «Потери инвалида».

  Тогда величиной, минимизирующей функцию $\mathbb{E}S(t, y(\boldsymbol{x}, \boldsymbol{w}))$, является следующая
  - $y(\boldsymbol{x}) = \mathbb{E}p(t|\boldsymbol{x})$;
  - $y(\boldsymbol{x}) = \text{med}\ p(t|\boldsymbol{x})$;
  - $y(\boldsymbol{x}) = \text{mod}\ p(t|\boldsymbol{x}) = \arg\max_t p(t|\boldsymbol{x})$.
- В зависимости от выбранной системы предпочтений, мы будем пытаться оценивать тот или иной функционал от апостериорного распределения вместо того, чтобы оценивать его самого

План лекции

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

Классическая линейная регрессия

Метод наименьших квадратов

Вероятностная постановка задачи

Задача классификации

Ликбез
- Нормальное распределение
- Решение нерешаемых СЛАУ

Статистическая постановка задачи машинного обучения
- Вероятностное описание
- Байесовские решающие правила

Линейная регрессия
- Классическая линейная регрессия
- Метод наименьших квадратов
- Вероятностная постановка задачи

Задача классификации
- Логистическая регрессия
- Метод IRLS

Минимизация невязки

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

Классическая линейная регрессия

Метод наименьших квадратов

Вероятностная постановка задачи

Задача классификации

- Наиболее часто используемой функцией потерь является квадратичная $S(t, \hat{t}) = (t - \hat{t})^2$
- Значение регрессионной функции на обучающей выборке в матричном виде может быть записано как $\boldsymbol{y} = \Phi \boldsymbol{w}$, где $\Phi = (\phi_{ij}) = (\phi_j(\boldsymbol{x}_i)) \in \mathbb{R}^{n \times m}$
- Таким образом, приходим к следующей задаче

$$\|\boldsymbol{y} - \boldsymbol{t}\|^2 = \|\Phi \boldsymbol{w} - \boldsymbol{t}\|^2 \to \min_{\boldsymbol{w}}$$

Взяв производную по $\boldsymbol{w}$ и приравняв ее к нулю, получаем

$$\frac{\partial \|\Phi \boldsymbol{w} - \boldsymbol{t}\|^2}{\partial \boldsymbol{w}} = \frac{\partial [\boldsymbol{w}^T \Phi^T \Phi \boldsymbol{w} - 2\boldsymbol{w}^T \Phi^T \boldsymbol{t} + \boldsymbol{t}^T \boldsymbol{t}]}{\partial \boldsymbol{w}} =$$

$$= 2\Phi^T \Phi \boldsymbol{w} - 2\Phi^T \boldsymbol{t} = 0$$

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

Регуляризация задачи

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

Классическая линейная регрессия

Метод наименьших квадратов

Вероятностная постановка задачи

Задача классификации

- Заметим, что формула для весов линейной регрессии представляет собой псевдорешение уравнения $\Phi \boldsymbol{w} = \boldsymbol{t}$
- Матрица $\Phi^T \Phi \in \mathbb{R}^{m \times m}$ вырождена (Упр.)при $m > n$
- Регуляризуя вырожденную матрицу, получаем

$$\boldsymbol{w} = \left(\Phi^T \Phi + \lambda I\right)^{-1} \Phi^T \boldsymbol{t}$$

- Отсюда формула для прогноза объектов обучающей выборки по их правильным значениям

$$\hat{\boldsymbol{t}} = \boldsymbol{y} = \Phi \left(\Phi^T \Phi + \lambda I\right)^{-1} \Phi^T \boldsymbol{t} = H\boldsymbol{t}$$

С историческим обозначением прогноза — навешиванием шляпки связано неформальное название матрицы H, по-английски звучащее как hat-matrix

Особенности квадратичной функции потерь

- Достоинства
  - Квадратичная функция потерь гладкая (непрерывная и дифференцируемая)
  - Решение может быть получено в явном виде
  - Существует простая вероятностная интерпретация прогноза и функции потерь
- Недостатки
  - Решение неустойчиво (не робастно) относительно даже малого количества выбросов. Это связано с быстрым возрастанием квадратичной функции потерь при больших отклонениях от нуля
  - Квадратичная функция неприменима к задачам классификации

План лекции

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

Классическая линейная регрессия

Метод наименьших квадратов

Вероятностная постановка задачи

Задача классификации

Ликбез
- Нормальное распределение
- Решение нерешаемых СЛАУ

Статистическая постановка задачи машинного обучения
- Вероятностное описание
- Байесовские решающие правила

Линейная регрессия
- Классическая линейная регрессия
- Метод наименьших квадратов
- Вероятностная постановка задачи

Задача классификации
- Логистическая регрессия
- Метод IRLS

Нормальное распределение ошибок

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

Классическая линейная регрессия

Метод наименьших квадратов

Вероятностная постановка задачи

Задача классификации

- Рассмотрим вероятностную постановку задачи восстановления регрессии. Регрессионная переменная t — случайная величина с плотностью распределения $p(t|\boldsymbol{x})$
- В большинстве случаев предполагается, что t распределена нормально относительно некоторого мат. ожидания $y(\boldsymbol{x})$, определяемого точкой $\boldsymbol{x}$

$$t = y(\boldsymbol{x}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\varepsilon|0, \sigma^2)$$

- Необходимо найти функцию $y(\boldsymbol{x})$, которую мы можем отождествить с уравнением регрессии
- Предположение о нормальном распределении отклонений можно обосновать ссылкой на центральную предельную теорему

Метод максимального правдоподобия для регрессии

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

Классическая линейная регрессия

Метод наименьших квадратов

Вероятностная постановка задачи

Задача классификации

- Используем ММП (не путать с одноименной кафедрой) для поиска $y(\boldsymbol{x})$
- Правдоподобие задается следующей формулой

$$p(\boldsymbol{t}|\boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t_i - y_i)^2}{2\sigma^2}\right) \to \max$$

- Взяв логарифм и отбросив члены, не влияющие на положение максимума, получим

$$\sum_{i=1}^{n}(t_i - y_i)^2 = \sum_{i=1}^{n}(t_i - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2 \to \min_{\boldsymbol{w}}$$

- Таким образом, применение метода максимального правдоподобия в предположении о нормальности отклонений эквивалентно методу наименьших квадратов

Вероятностный смысл регуляризации

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

Классическая линейная регрессия

Метод наименьших квадратов

Вероятностная постановка задачи

Задача классификации

- Теперь будем максимизировать не правдоподобие, а апостериорную вероятность
- По формуле условной вероятности

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

знаменатель не зависит от $\boldsymbol{w}$, поэтому им можно пренебречь

- Пусть $p(\boldsymbol{w}) \sim \mathcal{N}\left(\boldsymbol{w} \middle| \boldsymbol{0}, \left(\frac{\sigma^2}{\lambda}\right) I\right)$. Тогда

$$p(\boldsymbol{w}|\boldsymbol{t}, X) \propto \frac{\lambda^{m/2}}{\left(\sqrt{2\pi}\sigma\right)^{m+n}} \exp\left(-\frac{1}{2}\left(\sigma^{-2}\|\Phi\boldsymbol{w} - \boldsymbol{t}\|^2 + \frac{\lambda}{\sigma^2}\|\boldsymbol{w}\|^2\right)\right)$$

- Логарифмируя и приравнивая производную по $\boldsymbol{w}$ к нулю, получаем

$$\boldsymbol{w} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T\boldsymbol{t}$$

- Регуляризация эквивалентна введению априорного распределения, поощряющего небольшие веса

Зачем нужна регугляризация весов

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

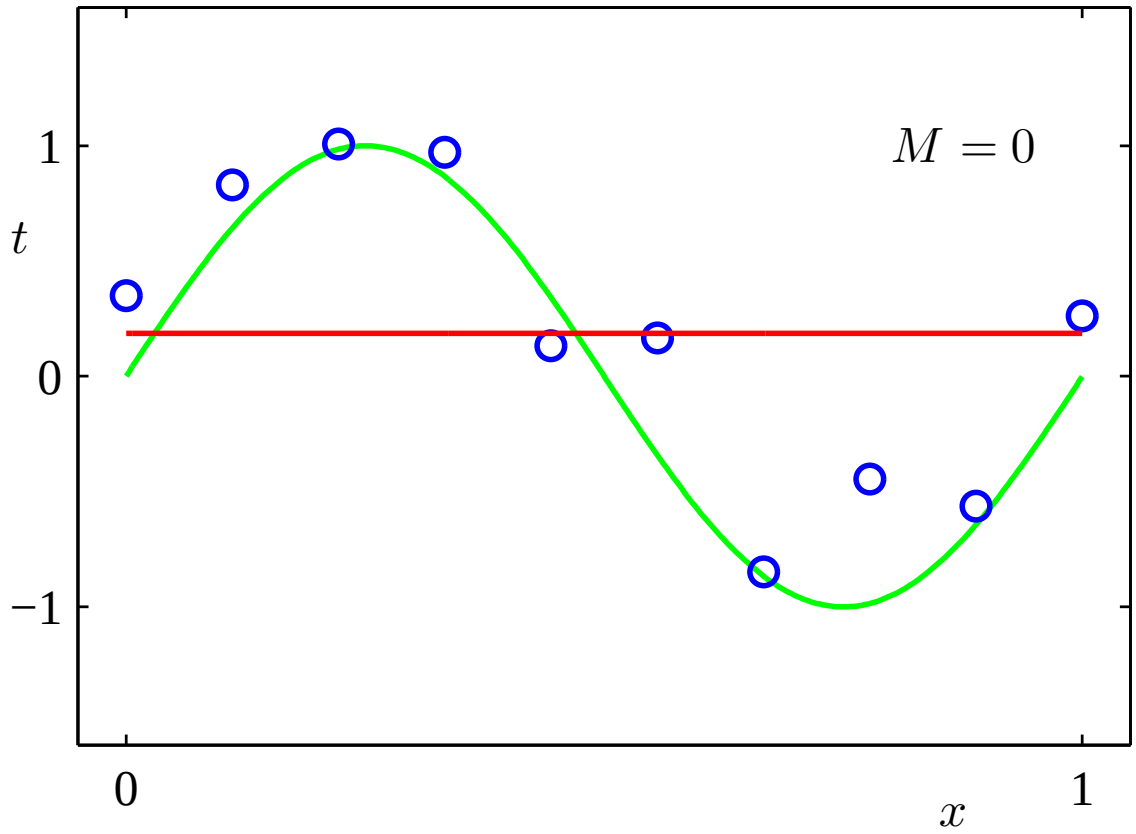
Классическая линейная регрессия

Метод наименьших квадратов

Вероятностная постановка задачи

Задача классификации

Рассмотрим задачу восстановления регрессии с полиномиальными базисными функциями: $x \in \mathbb{R}$, $\phi_j(x) = x^j$, $j = 0, \ldots, M$



Зачем нужна регугляризация весов

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

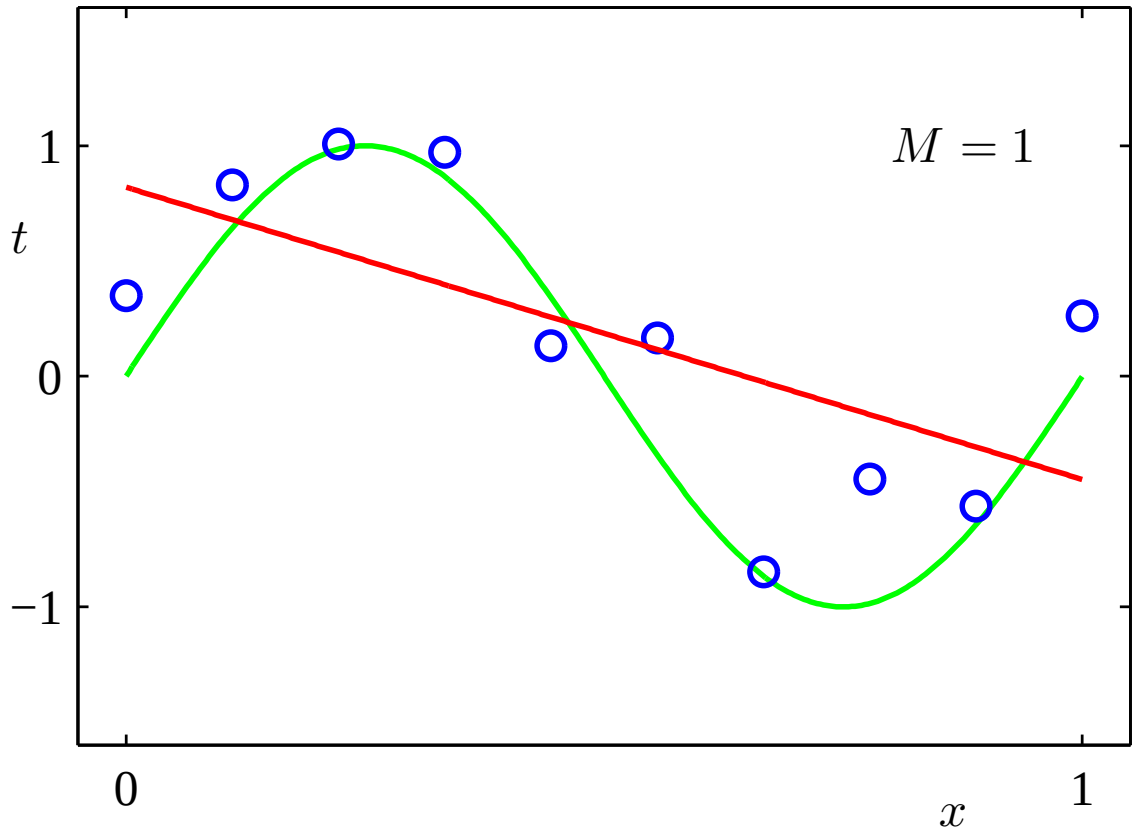
Классическая линейная регрессия

Метод наименьших квадратов

Вероятностная постановка задачи

Задача классификации

Рассмотрим задачу восстановления регрессии с полиномиальными базисными функциями: $x \in \mathbb{R}$, $\phi_j(x) = x^j$, $j = 0, \ldots, M$



Зачем нужна регугляризация весов

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

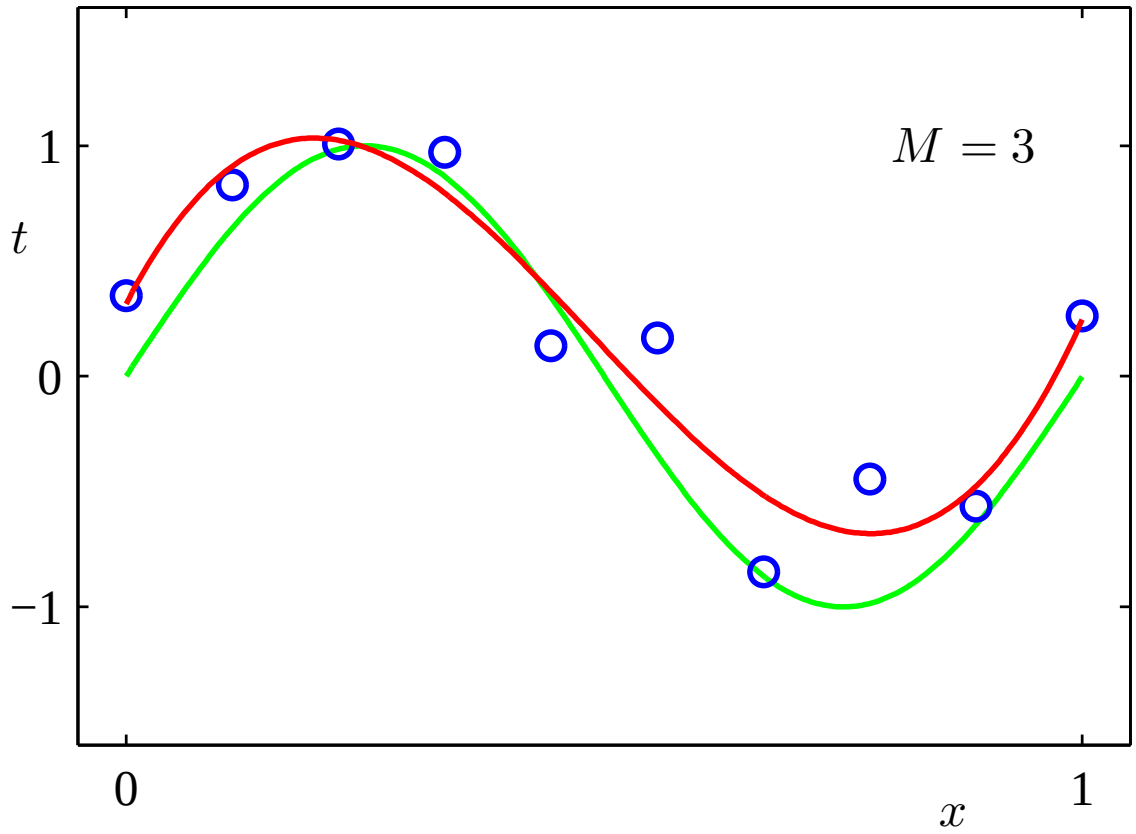
Классическая линейная регрессия

Метод наименьших квадратов

Вероятностная постановка задачи

Задача классификации

Рассмотрим задачу восстановления регрессии с полиномиальными базисными функциями: $x \in \mathbb{R}$, $\phi_j(x) = x^j$, $j = 0, \ldots, M$



Зачем нужна регугляризация весов

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

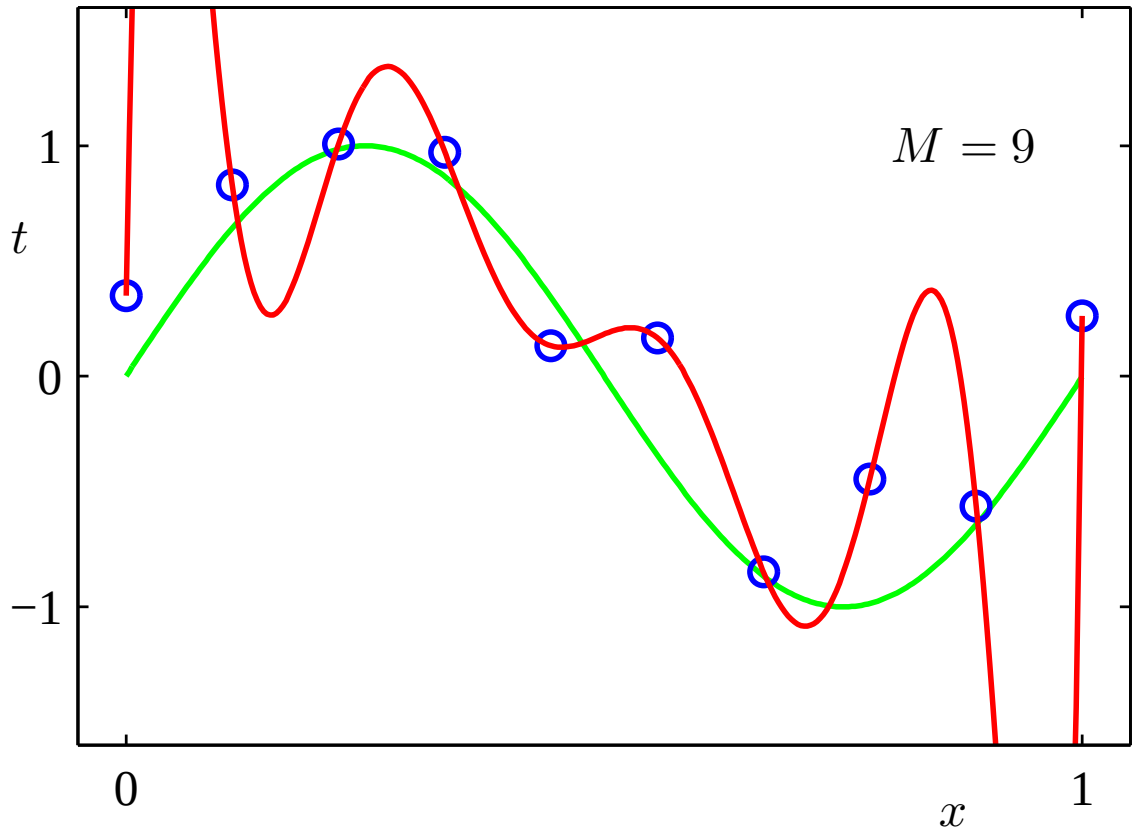
Классическая линейная регрессия

Метод наименьших квадратов

Вероятностная постановка задачи

Задача классификации

Рассмотрим задачу восстановления регрессии с полиномиальными базисными функциями: $x \in \mathbb{R}$, $\phi_j(x) = x^j$, $j = 0, \ldots, M$



Значения наиболее правдоподобных весов

weight	$M=0$	$M=1$	$M=3$	$M=9$
w_0	0.19	0.82	0.31	0.35
w_1		-1.27	7.99	232.37
w_2			-25.43	-5321.83
w_3			17.37	48568.31
w_4				-231639.30
w_5				640042.26
w_6				-1061800.52
w_7				1042400.18
w_8				-557682.99
w_9				125201.43

Таблица: Значения наиболее правдоподобных весов в зависимости от степени полинома. С увеличением степени, абсолютные значения весов быстро растут

План лекции

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

Задача классификации

Логистическая регрессия

Метод IRLS

Ликбез
- Нормальное распределение
- Решение нерешаемых СЛАУ

Статистическая постановка задачи машинного обучения
- Вероятностное описание
- Байесовские решающие правила

Линейная регрессия
- Классическая линейная регрессия
- Метод наименьших квадратов
- Вероятностная постановка задачи

Задача классификации
- Логистическая регрессия
- Метод IRLS

Особенности задачи классификации

- Рассмотрим задачу классификации на два класса $t \in \{-1, +1\}$
- Ее можно свести к задаче регрессии, например, следующим образом

$$\hat{t}(\boldsymbol{x}) = \text{sign}(y(\boldsymbol{x})) = \text{sign} \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x})$$

- Возникает вопрос: что использовать в качестве значений регрессионной переменной на этапе обучения?
- Наиболее распространенный подход заключается в использовании значения $+\infty$ для $t = +1$ и $-\infty$ для $t = -1$
- Геометрический смысл: чем дальше от нуля значение $y(\boldsymbol{x})$, тем увереннее мы в классификации объекта $\boldsymbol{x}$

Правдоподобие правильной классификации

- Метод наименьших квадратов, очевидно, неприменим при таком подходе
- Воспользуемся вероятностной постановкой для выписывания функционала качества
- Определим правдоподобие классификации следующим образом

$$p(t|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1 + \exp(-ty(\boldsymbol{x}))}$$

- Это логистическая функция. Легко показать, что $\sum_t p(t|\boldsymbol{x}, \boldsymbol{w}) = 1$ и $p(t|\boldsymbol{x}, \boldsymbol{w}) > 0$, а, значит, она является функцией правдоподобия

Функционал качества в логистической регрессии

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

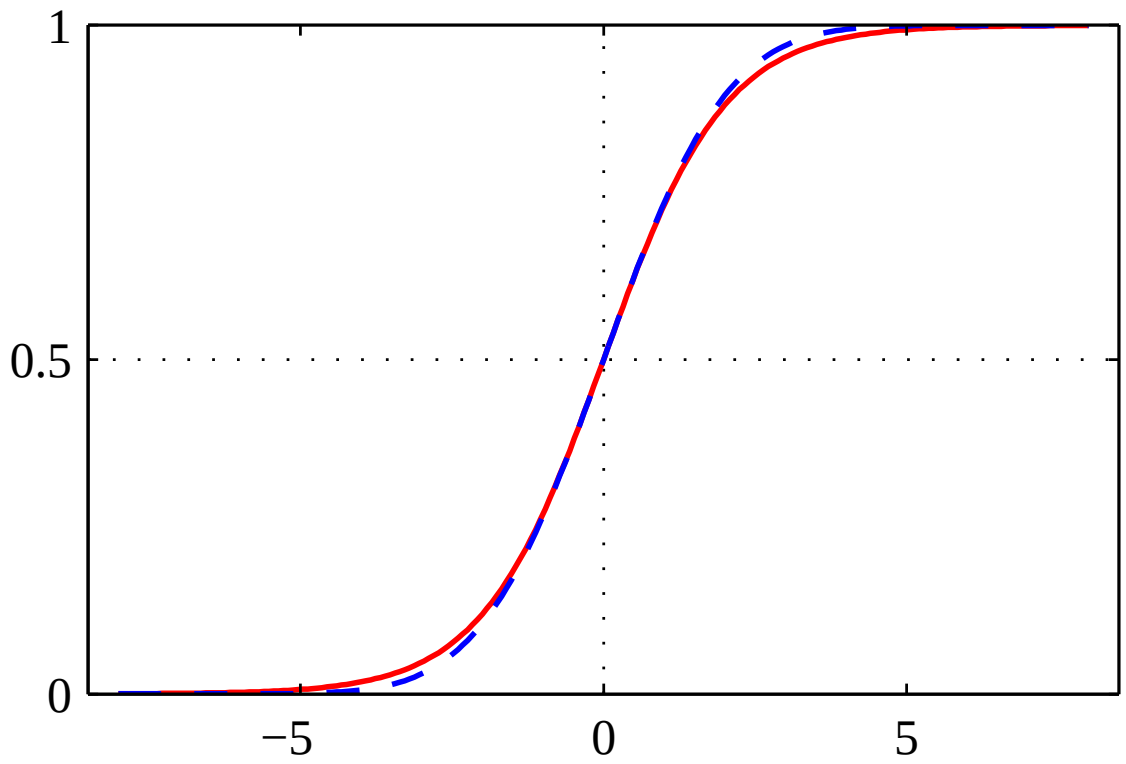
Статистическая постановка задачи машинного обучения

Линейная регрессия

Задача классификации

Логистическая регрессия

Метод IRLS



- Правдоподобие правильной классификации всей выборки имеет вид

$$p(\boldsymbol{t}|X,\boldsymbol{w}) = \prod_{i=1}^{n} p(t_i|\boldsymbol{x}_i,\boldsymbol{w}) = \prod_{i=1}^{n} \frac{1}{1+\exp\left(-t_i \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}_i)\right)}$$

План лекции

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

Задача классификации

Логистическая регрессия

Метод IRLS

Ликбез
- Нормальное распределение
- Решение нерешаемых СЛАУ

Статистическая постановка задачи машинного обучения
- Вероятностное описание
- Байесовские решающие правила

Линейная регрессия
- Классическая линейная регрессия
- Метод наименьших квадратов
- Вероятностная постановка задачи

Задача классификации
- Логистическая регрессия
- Метод IRLS

Особенности функции правдоподобия классификации

- Приравнивание градиента логарифма правдоподобия к нулю приводит к трансцендентным уравнениям, которые неразрешимы аналитически
- Легко показать, что гессиан логарифма правдоподобия неположительно определен

$$\frac{\partial^2 \log p(t|\boldsymbol{x}, \boldsymbol{w})}{\partial \boldsymbol{w}^2} \leq 0$$

- Это означает, что логарифм функции правдоподобия является вогнутым.
- Логарифм правдоподобия обучающей выборки $L(\boldsymbol{w}) = \log p(\boldsymbol{t}|X, \boldsymbol{w})$, являющийся суммой вогнутых функций, также вогнут, а, значит, имеет единственный максимум

Метод оптимизации Ньютона

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

Задача классификации

Логистическая регрессия

Метод IRLS

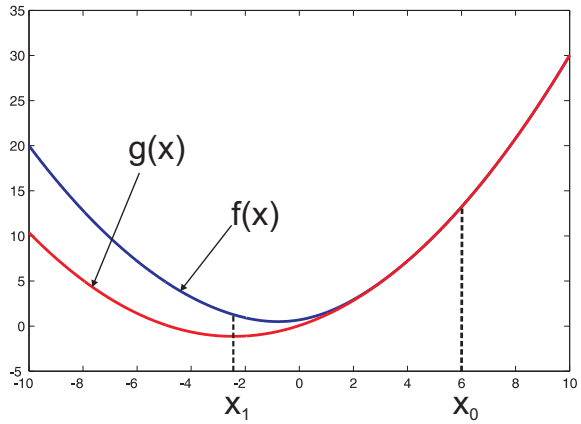
Основная идея метода Ньютона — это приближение в заданной точке оптимизируемой функции параболой и выбор минимума этой параболы в качестве следующей точки итерационного процесса:

$$f(\boldsymbol{x}) \to \min_{\boldsymbol{w}}$$

$$f(\boldsymbol{x}) \simeq g(\boldsymbol{x}) = f(\boldsymbol{x}_0) + (\nabla f(\boldsymbol{x}_0))^T(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T(\nabla\nabla f(\boldsymbol{x}_0))(\boldsymbol{x} - \boldsymbol{x}_0)$$

$$\nabla g(\boldsymbol{x}_*) = \nabla f(\boldsymbol{x}_0) + (\nabla\nabla f(\boldsymbol{x}_0))(\boldsymbol{x}_* - \boldsymbol{x}_0) = 0 \Rightarrow \boldsymbol{x}_* = \boldsymbol{x}_0 - (\nabla\nabla f(\boldsymbol{x}_0))^{-1}(\nabla f(\boldsymbol{x}_0))$$

Пример. Функция $f(x) = \log(1 + \exp(x)) + \frac{x^2}{5}$.
$x_0 = 6,\ x_1 = -2.4418.$



Итеративная минимизация логарифма правдоподобия

- Так как прямая минимизация правдоподобия невозможна, воспользуемся итерационным методом Ньютона
- Обоснованием корректности использования метода Ньютона является унимодальность оптимизируемой функции $L(\boldsymbol{w})$ и ее гладкость во всем пространстве весов
- Формула пересчета в методе Ньютона

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

где $H = \nabla\nabla L(\boldsymbol{w})$ — гессиан логарифма правдоподобия обучающей выборки

Формулы пересчета

Лекция 2. Вероятностная постановка задачи распознавания образов. Обобщенные линейные модели

Ветров, Кропотов

Ликбез

Статистическая постановка задачи машинного обучения

Линейная регрессия

Задача классификации

Логистическая регрессия

Метод IRLS

Обозначим $s_i = \frac{1}{1+\exp(-t_i y_i)}$, тогда:

$$\nabla L(\boldsymbol{w}) = \Phi^T(\boldsymbol{s} - \boldsymbol{t}), \quad \nabla\nabla L(\boldsymbol{w}) = \Phi^T R \Phi$$

$$R = \begin{pmatrix} s_1(1-s_1) & 0 & \dots & 0 \\ 0 & s_2(1-s_2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & s_n(1-s_n) \end{pmatrix}$$

$$\boldsymbol{w}^{new} = \boldsymbol{w}^{old} - (\Phi^T R \Phi)^{-1} \Phi^T (\boldsymbol{s} - \boldsymbol{t}) =$$
$$(\Phi^T R \Phi)^{-1} \left(\Phi^T R \Phi \boldsymbol{w}^{old} - \Phi^T R R^{-1} (\boldsymbol{s} - \boldsymbol{t}) \right) = (\Phi^T R \Phi)^{-1} \Phi^T R \boldsymbol{z},$$

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

Название метода (метод наименьших квадратов с итеративно пересчитываемыми весами) связано с тем, что последняя формула является формулой для взвешенного МНК (веса задаются диагональной матрицей R), причем на каждой итерации веса корректируются

Заключительные замечания

- На практике матрица $\Phi^T R \Phi$ часто бывает вырождена (всегда при $m > n$), поэтому обычно прибегают к регуляризации матрицы $(\Phi^T R \Phi + \lambda I)$
- !! Параметр регуляризации λ является структурным параметром!!
- !! Базисные функции $\phi_j(\boldsymbol{x})$, а значит и матрица Φ являются структурными параметрами!!
- С поиском методов автоматического выбора базисных функций связана одна из наиболее интригующих проблем современного машинного обучения