# Machine Learning

## Section 8: Linear Regression线性回归

Stefan Harmeling

3. November 2021

# Overview概述

- Regression
- Linear regression
- Maximum likelihood estimation
- Ridge regression里奇回归
- Bayesian linear regression贝叶斯线性回归
- Alternatives

This lecture is based on Chapter 7 of Kevin Murphy's textbook "Machine Learning, A Probabilistic Perspective"

# Regression (1)

数据点（x1，y1），. . . , (xn, yn)
xi是位置，通常是向量，有时是标量
yi是数值，通常是标量
目标：找到一个能将位置映射到数值上的函数f

## Setup

- data points $(x_1, y_1), \ldots, (x_n, y_n)$
- $x_i$ are locations, usually vectors, sometimes scalars
- $y_i$ are values, usually scalars
- goal: find a function $f$ that maps locations onto values

## Applications / why useful?

- predict celestial orbits (done by 24 year old Gauss, Ceres)
- interpolate measurements (e.g. between climate station)
- smooth noisy measurements (e.g. spectroscopy)
- predict the future (for time locations $x$)

预测天体轨道（由24岁的高斯完成，Ceres）。
插值测量（如气候站之间）。
平滑噪声测量（如光谱学）。
预测未来（对于时间地点X）。

# Regression (2)

**Procedure** ▸ assume a model for function $f$, e.g. for scalar $x$:假设一个函数f的模型，例如标量x的模型。

$$f(x) = w_0 + w_1 x \qquad \text{linear in } x$$

$$f(x) = w_0 + w_1 x + w_2 x^2 \qquad \text{nonlinear in } x \text{非线性}$$

$$f(x) = w_0 + w_1 x + \ldots + w_d x^d \qquad \text{nonlinear in } x$$

- called linear regression since linear in parameter $w$
- fit model with least squares, i.e.

称为线性回归，因为参数w是线性的

用最小二乘法拟合模型，即

$$w_{\mathrm{LS}} = \arg\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

为什么是最小二乘法？
为什么不是绝对值？
最小二乘法背后的假设是什么？

**Questions**

- why least squares? why not absolute values?
- what are the assumptions behind least squares?

# Some of the origins of method of least squares
# 最小二乘法的一些起源

- C.F. Gauss. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum*, 1809.
  - method of least squares
  - method of maximum likelihood
  - method of normal distribution
- A.M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*, 1805.

Source:

- `http://en.wikipedia.org/wiki/Regression_analysis#History`
- `http://en.wikipedia.org/wiki/Normal_distribution#Development`

# Notation with many variations
# 有许多变化的记号

# Linear regression: model specification (1)
# 线性回归：模型规格 (1)

## Single data point / linear function单一数据点/线性函数

- location $x$ is a vector位置x是一个向量
- function value at $x$ is modelled as $x^T w$ which is linear in $x$
- *measured* value $y$ is Gaussian distributed around $x^T w$

$$p(y|x, w) = \mathcal{N}(y|x^T w, \sigma^2) \qquad \text{univariate}$$

- $\sigma^2$ is the variance of the measurement noise方差
- value $y$ is scalar标量
- parameter $w$ is unknown未知的
- parameter $\sigma^2$ is known已知的
- because $x^T w$ is linear in $w$ this is linear regression

# Linear regression: model specification (2)

## Multiple data points / linear function多个数据点/线性函数

- location matrix $X$ contains vectors $x_1, \ldots, x_n$ as rows

位置矩阵X包含向量x1, . . . , xn为行

(why rows? see next point)

- function values at $X$ are modelled as $Xw$ which is linear in $X$ (using rows in $X$ makes $Xw$ really simple and minimalistic)
- *measured* values $y$ are Gaussian distributed around $Xw$

$$p(y|X, w) = \mathcal{N}(y|Xw, \Sigma) \qquad \text{multivariate}$$

- vector $y$ contains for each $x_i$ a scalar value
- because $Xw$ is linear in $w$ this is linear regression

## Towards linear regression for nonlinear functions
## 迈向非线性函数的线性回归

# Basis function expansion基准函数扩展

- for scalar $x$ the polynomial basis function对标量x的多项式基函数

$$\phi(x) = [1, x, x^2, \ldots, x^d]^T$$

leads to polynomials in $x$

$$\phi(x)^T w = \sum_{i=0}^{d} w_i x^i = w_0 + w_1 x + w_2 x^2 + \ldots + w_d x^d$$

- for vector $x = [x_1, x_2]$ the polynomial basis function

$$\phi(x) = [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2, \ldots, x_1^d, x_2^d]^T$$

leads to polynomials in $x_1$ and $x_2$ (or simply in $x$):

$$\phi(x)^T w = \sum_{i+j\leq d} w_{ij} x_1^i x_2^j$$

$$= w_{00} + w_{10} x_1 + w_{01} x_2 + w_{20} x_1^2 + w_{02} x_2^2 + w_{11} x_1 x_2 + \ldots + w_{d0} x_1^d + w_{0d} x_2^d$$

- in general $\phi$ maps vector $x$ nonlinearly onto vector $\phi(x)$
- the entries of vector $\phi(x)$ are also called *features*
- $\phi(x)^T w$ is linear in $w$ and possibly nonlinear in $x$

# Linear regression: model specification (3)

Single data point / nonlinear function单一数据点/非线性函数

- function value at $x$ is modelled as $\phi(x)^T w$ which is nonlinear in $x$
- *measured* value $y$ is Gaussian distributed around $\phi(x)^T w$

$$p(y|x,w) = \mathcal{N}(y|\phi(x)^T w, \sigma^2) \qquad \text{univariate}$$

- because $\phi(x)^T w$ is linear in $w$ this is linear regression

# Linear regression: model specification (4)

## Multiple data points / nonlinear function

- $\phi(X)$ is the matrix with rows $\phi(x_1), \ldots, \phi(x_n)$
- function values are modelled as $\phi(X)w$ which is nonlinear in $X$
- *measured* values $y$ are Gaussian distributed around $\phi(X)w$

$$p(y|X, w) = \mathcal{N}(y|\phi(X)w, \Sigma) \qquad \text{multivariate}$$

- because $\phi(X)w$ is linear in $w$ this is linear regression

## Notes

- $\phi(X)$ has just new locations along the rows
- for readability we consider only $X$ instead of $\phi(X)$
- however, all results hold for both $X$ and $\phi(X)$

Goal: estimate parameter vector $w$

Question

*Why is linear regression called linear?*

Answers:

- A Because it is linear in the features.
- B Because it is linear in the parameters.
- C Because it honors Francois Philippe Marquis de l'Inéar.
- D Because it sounds more scientific than just regression.

# Remember:

Linear regression is linear
because it is linear in the parameters.
线性回归是线性的
因为它在参数上是线性的。

# Maximum likelihood estimation (1)

## ML estimator

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

- iid data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- “iid” means *independent identically distributed* "iid "指独立同分布
- iid implies that likelihood factorizes iid 意味着可能性因素化

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{n} p(y_i|x_i, \theta)$$

- this (common) notation is a bit weird, $\mathcal{D}$ was only left of bar, but on RHS $x_i$ is conditioned on
- however, that’s ok, the location is always assumed to be known, also for prediction
- so more precisely, $\mathcal{D}$ only contains the values $y_1, \ldots, y_n$
- better: $p(y|X, \theta) = \ldots$

# Maximum likelihood estimation (2)

Instead to look at the likelihood we consider the

Log-likelihood

$$
\begin{aligned}
\ell(w) &= \log p(\mathcal{D}|w) = \sum_{i=1}^{n} \log p(y_i|x_i, w) \\
&= \sum_{i=1}^{n} \log \mathcal{N}(y_i|x_i^T w, \sigma^2) \\
&= -\frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{n} (y_i - x_i^T w)^2}_{\text{mean squared error}} - \frac{n}{2} \log(2\pi\sigma^2)
\end{aligned}
$$

- mean squared error (MSE) is also called *sum of squared error*, $\ell_2$ norm of residual errors, etc.
- ML estimation assuming a Gaussian likelihood leads to the method of least squares

# Maximum likelihood estimation (3)

假设高斯可能性的ML估计导致了最小二乘法的出现
因此：如果我们有高斯分布的测量值，那么最小二乘法是一个很合理的方法（通过ML）。
在高斯1809年的论文中，他从平均数开始，而平均数是科学中既定的估计指标（因为它是直观的），并想知道使用平均数意味着什么分布。由此，他发明了正态分布。

- ML estimation assuming a Gaussian likelihood leads to the method of least squares
- Thus: if we have Gaussian distributed measurements, then least squares is a well-justified method (via ML)
- In Gauss' paper from 1809, he started with the mean which was an established estimator in science (since it is intuitive) and wondered what distribution implies using the mean. By this he invented the normal distriution.

# Maximum likelihood estimation (4)

Likelihood

$$p(y|X, w) = \mathcal{N}(y|Xw, \Sigma)$$

Closed-form solution for the ML estimator

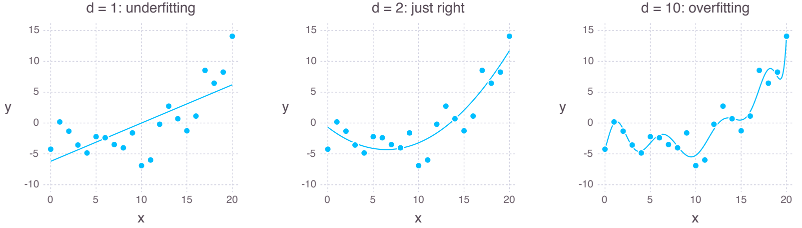
$$w_{\mathrm{ML}} = (X^T X)^{-1} X^T y$$

- aka *ordinary least squares* (OLS)
- OLS can be derived by setting the derivative of $\log \mathcal{N}(y|Xw, \Sigma)$ wrt $w$ to zero and solve for $w$

OLS可以通过将log N
(ySXw, Σ)相对于w的导数设
为零并求解w而得到。

# Maximum likelihood estimation (5)

| $d = 1$ | $d = 2$ | $d = 10$ |
|---|---|---|
| $f(x) = w_0 + w_1 x$ | $f(x) = w_0 + w_1 x + w_2 x^2$ | $f(x) = \sum_{i=0}^{10} w_i x^i$ |




Notes
- *underfitting* happens if the model is not flexible enough

如果模型不够灵活，就会发生欠拟合。
- *overfitting* happens if the model is too flexible for the amount of data 过度拟合是指模型对所需数量来说过于灵活。

# Ridge regression (1)

## Overfitting of ML

- too many parameter, too little data参数太多，数据太少 ▸

usually the weights are very large 通常权重非常大

## Idea

- encourage smoother solutions by putting a zero-mean Gaussian prior on $w$ to keep it small

通过在w上设置零均值的高斯先验来鼓励更平滑的解决方案，以保持它的小。

$$p(w) = \mathcal{N}(w|0, \tau^2 I)$$

- the variance $\tau^2$ controls the strength of this prior ▸

do MAP estimation

方差 τ 2控制这个先验的强度。

做MAP估计

# Ridge regression (2)

## MAP estimation

$$\begin{aligned}
w_{\text{ridge}} &= \operatorname{argmax}_w p(w|X, y) = \operatorname{argmax}_w p(y|X, w)p(w|X)/p(y|X) \\
&= \operatorname{argmax}_w p(y|X, w)p(w) \\
&= \operatorname{argmax}_w \sum_{i=1}^{n} \log \mathcal{N}(y_i|x_i^T w, \sigma^2) + \sum_{j=1}^{d} \log \mathcal{N}(w_j|0, \tau^2) \\
&= \operatorname{argmin}_w \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T w)^2}_{\text{fit}} + \underbrace{\lambda \|w\|_2^2}_{\text{regularizer}}
\end{aligned}$$

with $\lambda = \sigma^2/\tau^2$ (just move the $\sigma^2$ from the first summand to the second summand and merge with $\tau^{-2}$) and $\|w\|_2^2 = \sum_j w_j^2$.

## Solution

$$w_{\text{ridge}} = (\lambda I + X^T X)^{-1} X^T y$$

# Ridge regression (3)

## Solution

$$w_{\text{ridge}} = \operatorname{argmin}_w \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T w)^2 + \lambda\|w\|_2^2$$
$$= (\lambda I + X^T X)^{-1} X^T y$$

## Notes

- ridge regression is the same as penalized least squares
- $\lambda\|w\|_2^2$ is $\ell_2$ regularization (aka weight decay)

# Ridge regression (4)

**Question**

- Why regularize by adjusting $\lambda$, why not changing $d$?

为什么要通过调整λ来规范化，为什么不改变d？

**Towards an answer**

- $d$ sets model complexity
- $\lambda$ measures inverse signal-to-noise ratio (next slides)

d设定模型的复杂性

λ测量反信噪比（下一张幻灯片）。

# Bayesian linear regression (1)贝叶斯线性回归

### Question

- Can we also derive the posterior distribution over $w$ (instead of point estimates via ML and MAP)?

我们是否也能得出w的后验分布（而不是通过ML和MAP的点估计）？

### Prior and likelihood先验和可能性

$$p(w) = \mathcal{N}(w|w_0, V_0)$$
$$p(y|X, w) = \mathcal{N}(y|Xw, \Sigma)$$

### Posterior后部

$$p(w|X, y) = \mathcal{N}(w|w_n, V_n)$$

$$V_n = (X^T\Sigma^{-1}X + V_0^{-1})^{-1}$$ posterior covariance后验协方差

$$w_n = V_n(V_0^{-1}w_0 + X^T\Sigma^{-1}y)$$ posterior mean后验平均数

# Bayesian linear regression (2)

## Posterior

$$p(w|X,y) = \mathcal{N}(w|w_n, V_n)$$

$$V_n = (X^T\Sigma^{-1}X + V_0^{-1})^{-1} \quad \text{posterior covariance后验协方差}$$

$$w_n = V_n(V_0^{-1}w_0 + X^T\Sigma^{-1}y) \quad \text{posterior mean后验平均数}$$

## Notes

- for $\Sigma = \sigma^2 I,\ V_0 = \tau^2 I,\ w_0 = 0$, the mean of the posterior corresponds to ridge regression后验的平均值对应于岭回归的结果

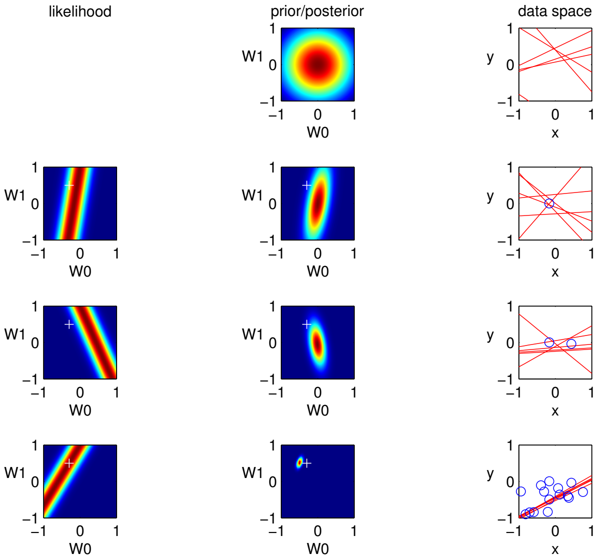
$$w_n = (\lambda I + X^T X)^{-1} X^T y = w_{\text{ridge}}$$

- however, here we have additionally the posterior covariance
然而，在这里我们还有后验协方差

$$V_n = \sigma^2(\sigma^2/\tau^2 I + X^T X)^{-1}$$
$$= \sigma^2(\lambda I + X^T X)^{-1}$$

  - so $\lambda$ is the inverse signal-to-noise ratio所以 $\lambda$ 是反信噪比

# Bayesian linear regression (3)




from Murphy's book Figure 7.11

# Bayesian linear regression (4)

## Posterior predictive distribution

- often we want to do prediction
- thus we integrate the parameter out
- training data $\mathcal{D}$ with $n$ pairs of locations and values
- how is value $y$ at a new data location $x$ distributed?

$$
\begin{aligned}
p(y|x,\mathcal{D}) &= \int \mathcal{N}(y|x^T w, \sigma^2)\,\mathcal{N}(w|w_n, V_n)\,dw \\
&= N(y|x^T w_n, \sigma_n^2) \\
\sigma_n^2 &= \sigma^2 + x^T V_n x
\end{aligned}
$$

- note that the variance is location dependent
- again for $\Sigma = \sigma^2 I$, $V_0 = \tau^2 I$, $w_0 = 0$:

$$
\sigma_n^2 = \sigma^2(1 + x^T(\lambda I + X^T X)^{-1} x)
$$

# Alternatives to least squares (1)

Linear regression

$$p(y|x, w) = \mathcal{N}(y|x^T w, \Sigma)$$

Robust linear regression

$$p(y|x, w) = \mathrm{Lap}(y|x^T w, b) = \exp(-\frac{1}{b}\|y - x^T w\|)/Z(b)$$

# Alternatives to least squares (2)

## Likelihoods and priors for linear regression

| Likelihood | Prior | Name |
|---|---|---|
| Gaussian | Uniform | Least squares |
| Gaussian | Gaussian | Ridge regression |
| Gaussian | Laplace | Lasso |
| Laplace | Uniform | Robust regression |
| Student | Uniform | Robust regression |

copied from Murphy's book Table 7.1

- note that, uniform prior leads to ML
- however, such a uniform prior can often not be normalized

# End of Section 08