LEARNING OBJ:
- you have choices
- most choices are solved w/ lin. algebra

Ch 8.1: Discrete l2 and Intro to Approximation Theory

Tuesday, October 21, 2025 2:16 PM

i.e. - Linear Regression
aka How to be a data scientist in 50 minutes

Context

We've seen interpolation, which has two major use cases:

Note: we index $i, \dots, m$ now instead of $0, 1, \dots, n$

① fit some data $\{(x_i, y_i)\}_{i=1}^{m}$ with a curve p

② approximate a function f with a curve p
(so we choose x_i, then set $y_i = f(x_i)$)

"Interpolation" means $p(x_i) = y_i$

Ch 8 is about Approximation Theory, with the same two scenarios!

① fit data. "discrete approximation" or "discrete least squares" or "l^2" ("little l 2")

② approximate a function, "continuous approximation" or "continuous least squares" or "L^2" ("big L 2")

...but instead of insisting $p(x_i) = y_i$, we only need $p(x_i) \approx y_i$.

Why not interpolate?

- If data are noisy, interpolation overfits to noise.
- Even if not noisy, interpolation is bad at extrapolation
- We might have physical / statistical reasons to keep p "simple" so it can't interpolate since not enough degrees-of-freedom.

Three types of categories:

① (1a) discrete $(x_1, x_2, \dots, x_m)$ or (1b) continuous $(\forall x \in [a,b] \dots)$

② what does $\underbrace{p(x_i)}_{\hat{y}_i} \approx y_i$ mean? Introduce a loss function E

As they come up, we'll discuss choices

③ What kind of function is p?

eg. polynomial ← different choices of basis
or something we can transform to a polynomial
or sines/cosines aka "trigonometric polynomials" } Fourier series!
any basis

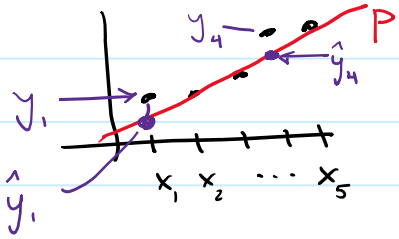
Tuesday, October 21, 2025 2:33 PM

Simplest scenario: discrete ℓ^2 "ordinary least squares" OLS

1. Discrete data $\{(x_i, y_i)\}_{i=1}^m$
2. loss is squared loss: $E = \sum_{i=1}^m (\hat{y}_i - y_i)^2$
3. curve p is a polynomial

3a. linear polynomial, $p(x) = a_0 + a_1 x$

So $\hat{y}_i = p(x_i)$ is our "prediction" or "estimate" of y_i



So...

$$\underset{a_0, a_1}{\text{minimize}} \quad \underbrace{\sum_{i=1}^m (a_0 + a_1 x_i - y_i)^2}_{E}$$

where $a_0 + a_1 x_i = \hat{y}_i$

We can solve w/ linear algebra! Minimizers must satisfy

$$\frac{\partial E}{\partial a_0} = 0 \quad \text{and} \quad \frac{\partial E}{\partial a_1} = 0.$$

In fact, since E is jointly convex in a_0 and a_1, these conditions are also sufficient.

$$(*) \quad \frac{\partial E}{\partial a_0} = \sum_{i=1}^m 2 \cdot (a_0 + a_1 x_i - y_i) = 0$$

$$(*) \quad \frac{\partial E}{\partial a_1} = \sum_{i=1}^m 2 x_i (a_0 + a_1 x_i - y_i) = 0$$

"normal" here refers to "orthogonal"

2 equations, linear in a_0 and a_1

Rewrite in vector form:

$$\vec{y} = (y_i) \in \mathbb{R}^m, \quad X = \begin{bmatrix} x_1^0 & x_1^1 \\ \vdots & \vdots \\ x_m^0 & x_m^1 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \in \mathbb{R}^{m \times 2}, \quad \vec{a} = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} \in \mathbb{R}^2$$

So $\hat{\vec{y}} = X \cdot \vec{a}$

and we want $\min_{\vec{a} \in \mathbb{R}^2} \| X\vec{a} - \vec{y} \|_2^2$ since $\|\vec{r}\|_2 = \sqrt{\sum_i r_i^2}$

$$= (X\vec{a} - \vec{y})^T (X\vec{a} - \vec{y}) = \vec{a}^T X^T X \vec{a} - 2\vec{y}^T X \vec{a} + \vec{y}^T \vec{y}$$

Solution $\vec{a}$ satisfies the normal equations

$$\underbrace{X^T X}_{2\times 2} \vec{a} = \underbrace{X^T \vec{y}}_{2 \times 1}$$

equivalent to (*)

Tatoo this to your arm. Don't forget it.

Tuesday, October 21, 2025 2:51 PM

Variations:

① non-linear polynomials, eg. $p(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$ (say, degree n)

then set $X = \begin{bmatrix} x_1^0 & x_1^1 & x_1^2 & x_1^3 \\ \vdots & & & \\ x_m^0 & x_m^1 & x_m^2 & x_m^3 \end{bmatrix} \in \mathbb{R}^{m \times n+1}$, $\vec{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} \in \mathbb{R}^{n+1}$

and solve $X^T X \vec{a} = X^T \vec{y}$ as before

Comments:

① Does $X^T X \vec{a} = X^T \vec{y}$ have a solution? Is it unique?

($X^T X$: Gram matrix)

If points x_i are distinct and $n+1 \leq m$ then columns of X are linearly independent, so $X^T X$ is positive definite ($\Rightarrow$ invertible), hence there is a solution and it's unique

② X is a Vandermonde type matrix and ill-conditioned if n is large.

Two tricks: ① Choose a better polynomial basis. More on this later this chapter

② Don't form $X^T X$ since that exacerbates the ill-conditioning. More on this in ch. 6 (eg., QR or SVD) Short answer: use `np.linalg.lstsq(X, y)`

② Change-of-variables

ex. $y = a_0 \cdot b^{a_1 x}$ or $y = a_0 x^{a_1}$

take $\log(\cdot)$ of these models: make it linear

$\log(y) = \log(a_0) + a_1 \cdot \log(b) \cdot x$ or $\log(y) = \log(a_0) + a_1 \cdot \log(x)$

rename: y, a_0, a_1, x

Solve for new variables, then transform back

⚠ The transformation affects how we interpret our loss

OR ... keep it nonlinear and solve non-linear least squares using ch. 10 techniques

Wednesday, October 22, 2025 9:42 AM

Other tricks

model: $y = \alpha_1 \cdot \sin(\omega t + \alpha_2)$

ω: known frequency

α_1: fit amplitude

α_2: ...and phase

Trick $\sin(u+v) = \sin(u)\cdot\cos(v) + \cos(u)\cdot\sin(v)$

so fit $y = \tilde{\alpha}_1 \cdot \sin(\omega t) + \tilde{\alpha}_2 \cdot \cos(\omega t)$ linear in parameters!

$$\left.\begin{aligned} \tilde{\alpha}_1 &= \alpha_1 \cdot \cos(\alpha_2) \\ \tilde{\alpha}_2 &= \alpha_1 \cdot \sin(\alpha_2) \end{aligned}\right\}$$

solve: $\tilde{\alpha}_1^2 + \tilde{\alpha}_2^2 = \alpha_1^2(\cos^2 + \sin^2) = \alpha_1^2$

to find α_1,

then $\alpha_2 = \cos^{-1}(\tilde{\alpha}_1/\alpha_1)$.

nonlinear: `scipy.optimize.least_squares`

linear: `scipy.linalg.lstsq` Don't use if you don't need to!

Tuesday, October 21, 2025 2:59 PM

③ Weighted & Generalized Least Squares

if $E = \sum_{i=1}^{m} w_i (\hat{y}_i - y_i)^2$ (eg. w_i reflects an estimate of accuracy of i^{th} data point)

then if $W = \text{diag}(\vec{w})$,

solve $\min_{\vec{a}} (X\vec{a} - \vec{y})^T \cdot W \cdot (X\vec{a} - \vec{y})$

w/ solution $X^T W X \vec{a} = X^T W \cdot \vec{y}$ Weighted least squares

even more generally, let $C > 0$ be positive definite,

solve $\min_{\vec{a}} (X\vec{a} - \vec{y})^T \cdot C \cdot (X\vec{a} - \vec{y})$ Ex: C = estimate of inverse of covariance matrix

often written $\|X\vec{a} - \vec{y}\|_C^2$ "Mahalanobis distance"

w/ solution $X^T C X \vec{a} = X^T C \vec{y}$ Generalized Least Squares

⚠ Careful w/ terminology.

"General linear model" is for multiple RHS $\vec{y}$

"Generalized linear model" if you have a "link" function, as in logistic regression.

Neither is the same as Generalized L.S.

④ (4a) $E = \sum_{i=1}^{m} |\hat{y}_i - y_i|$ "least absolute deviation" aka $\min_{\vec{a}} \|X\vec{a} - \vec{y}\|_1$

(4b) $E = \max_{1 \le i \le m} |\hat{y}_i - y_i|$ "minimax" aka $\min_{\vec{a}} \|X\vec{a} - \vec{y}\|_\infty$

Both reasonable, depend on where you think the "error" in data comes from. We'll explore in lab. Both are a special type of convex optimization problem known as a linear program. We've known how to solve since 1940's ... but Gauss, etc., couldn't solve. It's not just linear algebra.

Wednesday, October 22, 2025 6:36 AM

One way these ℓ_1 and ℓ_∞ error terms arise is in **Maximum likelihood estimation (MLE)**, a very common statistical technique.

MODEL: $y = \hat{y}_\theta + \text{noise}$

- y: we'll observe
- $\hat{y}_\theta$: our model, parameterized by θ
  - Ex: $\hat{y}_\theta = \theta_0 + \theta_1 x + \theta_2 x^2$
- noise: unobserved but we might know its distribution

Ex: assume noise $\sim N(0, \sigma^2)$ "normal" / "Gaussian"

so the probability density function (pdf) is $p_\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\hat{y}_\theta)^2/2\sigma^2}$

If we have multiple observations $\{y_i\}$ that are independent, then $p_\theta(\{y_i\}) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(y_i - \hat{y}_{i\theta})^2/2\sigma^2\right)$

abbreviate in lecture

The "likelihood" is just $p_\theta(\{y_i\}_{i=1}^m)$ but as a function of θ

Given a value of θ, it is "how likely would it be to see these observations if θ were the model parameters"

So, find the θ to maximize the likelihood... hence MLE.

Note: $p_\theta(\ldots) \geq 0$ so we could take $\log(p_\theta(\ldots))$ w/o changing location of maximizer. (log: monotonic)

So, $\max_\theta p_\theta(\{y_i\}) = \min_\theta -\log(p_\theta(\{y_i\}))$ (negative log-likelihood)

if noise $\overset{iid}{\sim} N(0,\sigma^2)$,

this is $\min_\theta -\log\left(\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(y_i - \hat{y}_{i\theta})^2/2\sigma^2\right)\right)$

i.e. $\min_\theta \sum_{i=1}^{m} (y_i - \hat{y}_{i\theta})^2 + \text{constants}$ (no θ dependence)

i.e. **least squares!** ← So popular because:

1. it's MLE for Gaussian noise, and Gaussians are common (ex: CLT)
2. it's easy to solve and well understood

Wednesday, October 22, 2025 9:29 AM

(5) Regularized least squares... in particular, Tikhonov / Ridge Regression

$$\min_{\vec{a}} \|X\vec{a} - \vec{y}\|_2^2 + \lambda \cdot \|\vec{a}\|_2^2$$

$\lambda > 0$ a hyperparameter

$\lambda \cdot \|\vec{a}\|_2^2$ ~ regularization term

Still has closed form solution:

$$(X^T X + \lambda I)\vec{a} = X^T \vec{y}$$

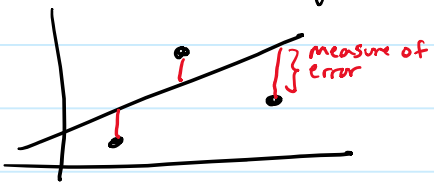
Many variants beyond $\|\vec{a}\|_2^2$

Arises in Maximum a posteriori (MAP) statistical estimation, which is a (kind of) Bayesian technique that assumes a "prior" distribution on $\vec{a}$ (not "fully" Bayesian since it's just a point estimate)

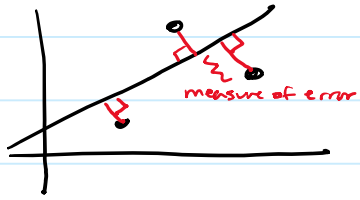
Tikhonov corresponds to assuming $\vec{a}$ is Gaussian.

(6) Total Least Squares & Orthogonal distance regression

Ordinary regression assumes we have error in our response variable y ... but what if you have error in your independent variable x ?



"Ordinary"



"total"

Can solve w/ SVD. We won't go into detail

Scipy.odr solves. You should provide it w/ an estimate of the variance in y and variance in x

if exactly equal, then error is exactly the perpendicular