KDE → small α → lots of bumps and less smooth

tabular [Excel sql
CSV/TSV: what if data has commas/tabs?
nested [XML / JSON: Each record can have diff fields
Records can contain records (nested)

Primary key: Unique ID for each entry
Foreign key: cols that ref primary keys in other tables

categorical data { Ordinal: has orders but no sense of magnitude/interval
nominal: no specific ordering

Quantitative: histograms, box plots, rug plots, smoothed interpolation (KDE): Look for spread, shape, modes, outliers, unreasonable values
nominal/ordinal: bar plots; look for skew, freq and rare categories, or invalid categories
consider grouping categories and repeating analysis
Inferential plot: draw conclusion beyond data

Gaussian kernel:

$$K_\alpha(r) = \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left(\frac{-r^2}{2\alpha^2}\right)$$

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} K_\alpha(x - x_i)$$

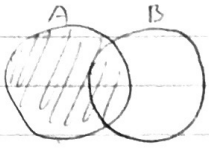
(query ↘, ↙ data)

Big n (many rows): aggregation & smoothing; avoid over-plotting or use transparency
Big p (many columns): use additional cols to adjust shape, size, color; combine cols
XML: elem must have open and close tag unless it is empty <tagname/>; must be properly nested, tag names case-sensitive, no space allowed btwn < and tag name, tag names must begin w/ letter & contain alphanum, attributes must be in quotes, use &lt for <, and &gt for >, must have one root node that contains all other nodes

Str methods: re.findall(pat, str) → series.str.findall
re.search(pat, str) replace, split, contains, len, [:] ← slicing

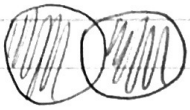
Rest API = Get, Post, PUT (update), delete (check (...) ↙)

SQL: schema: desc of cols, types and constraints
- instance: data satisfying the schema
- attribute (col), Tuple (record, row)
- DELETE from.
- UPDATE table SET gpa = 1.0 + gpa
- sorting: Order by gpa DESC, name ASC
- Need to use WHERE rating IS NOT NULL to find null vals or IS NULL
- Can use boolean logic in where clause

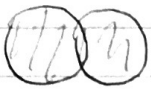


SELECT col FROM TableA A
LEFT JOIN TableB B
ON A.key = B.key

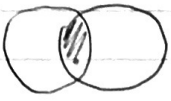
SELECT col FROM TableA A
LEFT JOIN TableB B
ON A.key = B.key
WHERE B.key IS NULL



" WHERE A.key IS NULL
OR B.key IS NULL



SELECT col from TableA A
FULL OUTER JOIN TableB B
ON A.key = B.key



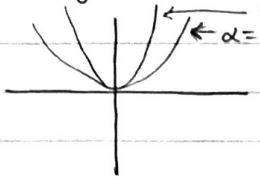
SELECT col FROM TableA A
INNER JOIN TableB B
ON A.key = B.key

Loss func: $L(\theta, y) = (y-\theta)^2$ L^2 loss
(predicted, observed)

$L(\theta, y) = |y-\theta|$ L^1 loss, abs loss

$$L_\alpha(\theta, y) = \begin{cases} \frac{1}{2}(y-\theta)^2 & |y-\theta| < \alpha \\ \alpha(|y-\theta| - \frac{\alpha}{2}) & \text{otherwise} \end{cases}$$

$\theta = y$ → good fit → no loss
θ far from y → bad fit → some loss



$\log_b M \cdot N = \log_b M + \log_b N$ $\quad \log_b b^k = k$

$\log_b(\frac{M}{N}) = \log_b M - \log_b N$ $\quad b^{\log_b(k)} = k$

$\log_b M^k = k \cdot \log_b M$ $\quad b>1$ & $M,N>0$

$\log_b(1) = 0$ $\quad \log_b b = 1$

HTTP Status
- 100s Informational
- 200s Success
- 300s Redirection or conditional action
- 400s Client Error
- 500s Internal Server Error or Broken Request

Average Loss

$L(\theta, D) = \frac{1}{n}\sum_{i=1}^{n} L(\theta, y_i)$

L^1 loss → less sensitive, not smooth

L^2 loss → sensitive, but smooth

Huber → less sensitive but extra param

second deriv has to be positive to be a min

minimize $L\theta = (1 - \log(1+\exp(\theta)))^2$

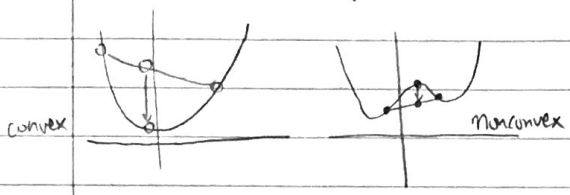
$\frac{\partial}{\partial\theta} L(\theta) = \frac{\partial}{\partial\theta}(1 - \log(1+\exp(\theta)))^2$

$= 2(1 - \log(1+\exp(\theta)))\frac{\partial}{\partial\theta}(1 - \log(1+\exp(\theta)))$

$= 2(1 - \log(1+\exp(\theta)))(-1)\frac{\partial}{\partial\theta}\log(1+\exp(\theta))$

$= 2(1 - \log(1+\exp(\theta)))\frac{-1}{1+\exp(\theta)}\frac{\partial}{\partial\theta}(1+\exp(\theta))$

$= 2(1 - \log(1+\exp(\theta)))\frac{-1}{1+\exp(\theta)}\exp(\theta)$



Derivatives

$f(x) = a^x \rightarrow f'(x) = \ln a \cdot a^x$

$t f(a) + (1-t) f(b) \geq f(ta + (1-t)b)\ \forall a, \forall b, t \in [0,1]$

$y = f(x)g(x) \rightarrow f'(x)g(x) + f(x)g'(x)$

Gradient Descent Algorithm $L(\theta, x, y) = \sum_{i=1}^{n} \theta^2 \cdot x_i^2 - \log(y_i)$

$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$

$\theta^{(t+1)} \leftarrow \theta^t - \alpha L'(\theta^t)$ $\quad L'(\theta) = \sum_{i=1}^{n} 2\theta x_i^2$

$\frac{dy}{dx} = \frac{dy}{du}\cdot\frac{du}{dx} = f'(u)\cdot g'(x)$

$= f'(g(x))\cdot g'(x)$

Init val θ^0 (zero, random guess)

for t until convergence update

α is the learning rate and not const ($\frac{1}{t}$)

might stop once $\nabla L(\theta^t)$ is small

Bayes: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

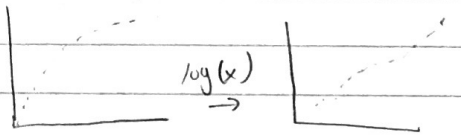
$= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$

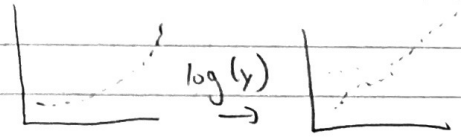
Data Collection sampling: SRS,

Stratified: SRS within partitions

cluster: divide pop into groups then take SRS of groups

Non Probability Samples: administrative, voluntary, convenience





EDA - Structure, granularity, scope, temporality, faithfulness

Pandas - .loc uses the value in the index col and not the actual row id → use iloc instead

Sort_values is ascending by default

df.['w'].value_counts()

sum(), min(), max(), count(), median(), var(), mean(), apply()

df.dot.hist() df.plot.scatter(x='w', y='h')

&, |, ~ df.loc[df['a'] > 10, ['a','c']]

rows satisfying boolean and only specific cols

SELECT FROM WHERE GROUP BY HAVING ORDER BY

[DISTINCT] for unique vals ASC|DESC

$\frac{\partial}{\partial\theta} L(\theta, D) = \sum_i (y_i - \theta x_i)^2$

$0 = \sum_i (-2x_i)(y_i - \theta x_i)$

$0 = \sum_i -2x_i y_i + 2\theta x_i^2$

$0 = -2\sum_i x_i y_i + 2\theta \sum_i x_i^2$

$\sum_i x_i y_i = \theta \sum_i x_i^2$

$\theta = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$

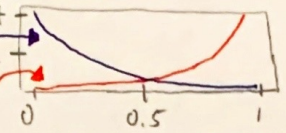
$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$

Average Cross Entropy Loss

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(\sigma(\phi(x_i)^T\theta)) + (1-y_i)\log(1-\sigma(\phi(x_i)^T\theta))\right)$$

If $y_i = 1 \Rightarrow -\log(\sigma(\phi(x_i)^T\theta))$

$y_i = 0 \Rightarrow -\log(1-\sigma(\phi(x_i)^T\theta))$ (prob $y_i = 0$)



log loss ⇒ log of the predicted prob for the true class

	Cute	Not cute
Observed Prob	$p(y=1\|x) = 1.0$	$p(y=0\|x) = 0$
Predicted Prob	$\hat{p}_\theta(y=1\|x) = 0.8$	$\hat{p}_\theta(y=0\|x) = 0.2$
Cross Entropy $-P \log \hat{P}_\theta$	$-1 \log(0.8) = 0.22$	$-0.0 \log(0.2) = 0.0$

Stochastic Gradient Descent

$$\nabla_\theta L(\theta) = \frac{1}{|B|}\sum_{i\in B}(\sigma(\phi(x_i)^T\theta) - y_i) * \phi(x_i)$$

$|B|$: Batch size; $i \in B$: random samples

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \rho(t)\left(\frac{1}{|B|}\sum_{i\in B}\nabla_\theta L_i(\theta)\Big|_{\theta=\theta^{(t)}}\right)$$

Linear models

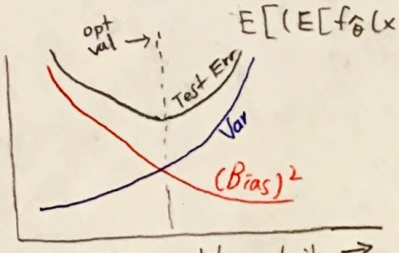
$f_\theta(x) = \theta_0 + \theta_1 + \theta_2 x^2$ (still linear in the parameters θ)

Bias: expected deviation between true and predicted value

Variance ⇒ observations ⇒ variability of random noise

⇒ model ⇒ variability in the predicted value across diff training sets

$$E[(y - f_\theta(x))^2] = \underbrace{E[(y-h(x))^2]}_{\text{noise}} + \underbrace{(h(x) - E[f_{\hat{\theta}}(x)])^2}_{\text{bias}^2} + \underbrace{E[(E[f_{\hat{\theta}}(x)] - f_{\hat{\theta}}(x))^2]}_{\text{Model Var}}$$



Least sq

$$\hat{y} = f_\theta(x) = \sum_{i=1}^{d}\theta_j \phi_j(x)$$

(linear in the params; $\phi_j(x)$: features)

$$\hat{\theta} = (\phi^T\phi)^{-1}\phi^T Y$$

$$0 = \phi^T(Y - \phi\theta)$$

Feature Eng

Quantitative → log, normalize

Categorical → one-hot-encode

Missing vals
- ↳ predict vals
- ↳ binary field for missing

Categorical
- ↳ True, False, missing
  1 −1 0

Bag of words
- ↳ text as long vect of word count

N-gram
- ↳ The book was well-written

Standardization
- ↳ each dimension has the same scale

Data Warehouse

Extracted from remote sources
Transformed to std schemas
Loaded into the relational data sys

Extract and Load - data in a single sys historical snapshot, isolates analytics

Transform - clean and prep data for analytics in a unified representation
- ↳ difficult b/c different schemas, encoding and granularities

Expectations

$$E[X] = \sum_{x\in X} x\,P(x)$$

$E[aX + Y + b] = aE[X] + E[Y] + b$

$E[XY] = E[X]E[Y]$ if X and Y indep.

$Var[X] = E[(X - E[X])^2]$

$= E[X^2] - E[X]^2$

$Var[aX+b] = a^2 Var[X] + 0$

$Var[X+Y] = Var[X] + Var[Y]$ if X, Y indep.

$SD[X] = \sqrt{Var[X]}$ $SD[aX+b] = |a| SD[X]$

Bernoulli

↳ $E[X] = p$ $Var[X] = p(1-p)$

Variance of sample mean decreases at rate $\frac{1}{n}$ ($\frac{\sigma^2}{n}$)

Standard Error $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Regularization

$$\hat{\theta} = \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n} Loss(y_i, f_\theta(x_i)) + \lambda R(\theta)$$

$$R_{Ridge}(\theta) = \sum_{i=1}^{d}\theta_i^2 \quad (L_2) \qquad R_{Lasso}(\theta) = \sum_{i=1}^{d}|\theta_i| \quad (L_1)$$

Ridge: distri weights across related features; analytical soln; small but nonzero weights

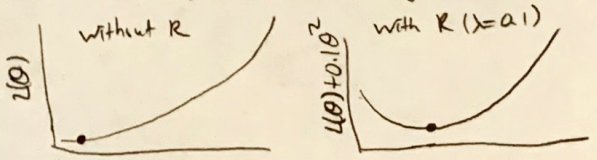
Lasso: encourage sparsity by setting weights to 0 ↳ used to select informative features; No analytical solution

Logistic Regression

$$\hat{P}_\theta(y=1|x) = \sigma(\underbrace{\phi(x)^T\theta}_{\text{linear model}}) = \frac{1}{1+\exp(-\phi(x)^T\theta)}$$

use cross entropy loss instead b/c non-convex

If linearly separable ⇒ needs regularization



$$P(Y=j|x) = \frac{\exp(x^T\theta^{(j)})}{\sum_{m=1}^{k}\exp(x^T\theta^{(m)})}$$

soft-max; multi-class

P-value

- measures the strength of evidence against null; smaller values, strong evidence
- can't turn rate problem into probability
- can't test if something is random using the assumption that it is random

Online Analytics Processing (OLAP)
- ↳ constructing complex SQL queries
- ↳ sharing views that summarize data across important dimensions

cross tabulation (pivot table)

Cube Operator
- ↳ generalizes cross-tabulation to higher dimensions.

Slicing ⇒ select val for a dimension

dicing ⇒ select a range of vals in multiple dimension

Rollup ⇒ Aggregate along a dimension

Drill-down ⇒ de-aggregating along a dimension

Data Lake
- ↳ store copy of all data in its natural form
- ↳ schema on read
- ↳ lots of dirty data
- ↳ hard to know what data contain and where data came from