Keywords

- what is ML
- test/train/validation
- linear algebra
- regularization
  - elastic net
  - $q=1$ LASSO $\sum e^2 + \alpha \frac{1}{2} \sum |\theta_i|$
  - $q=2$ Ridge / Tikhonov $\ldots + \alpha \frac{1}{2} \sum |\theta_i|^2$
  - $q=\infty$ minmax
- $x \in \mathbb{R}$
  $\min \|x\underline{1} - \underline{b}\|_q$ { median, mean, midrange
- missing data
- examples
- norms
- maximum likelihood
- feature engineering
- bayesian linear regression
- logistic regression (cross-entropy)
  $y = f(\underline{w}^T \underline{\phi})$
  ↑ activation function
  f^{-1} = link function

Part 1

- what is ML (tabular)
- examples, relational data
- linear regression
- norms
- feature engineering
- test/train/validation
- numpy, pandas, mpl, sklearn

Part 2

- Recap
- Regularization
- categorical data, missing data
- generalized linear models (poisson, logistic regression)
- sklearn pipelines
- tree models

Sources:

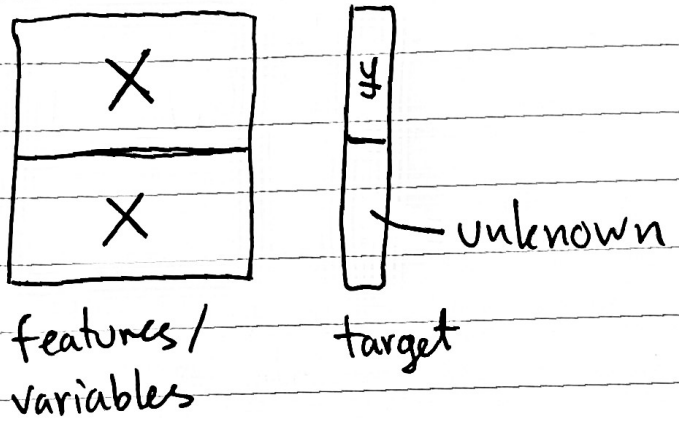
kap 6 "Approximation and fitting" boyd
kap 1,2,3,4,6,7 "Hands on ML" Géron
kap 3,4 "Pattern Recognition" Bishop
kap 3,4,9,10 "Elements" Hastie

Fagdag Sonat

What is machine learning?

$$X = \underset{n}{\begin{pmatrix} \text{---} & \underset{i}{\underline{x}_n} & \text{---} \end{pmatrix}} \qquad y = \underset{n}{\begin{pmatrix} y_n \end{pmatrix}}$$

- Pairs $(\underline{x}_n, y_n)$ in a table

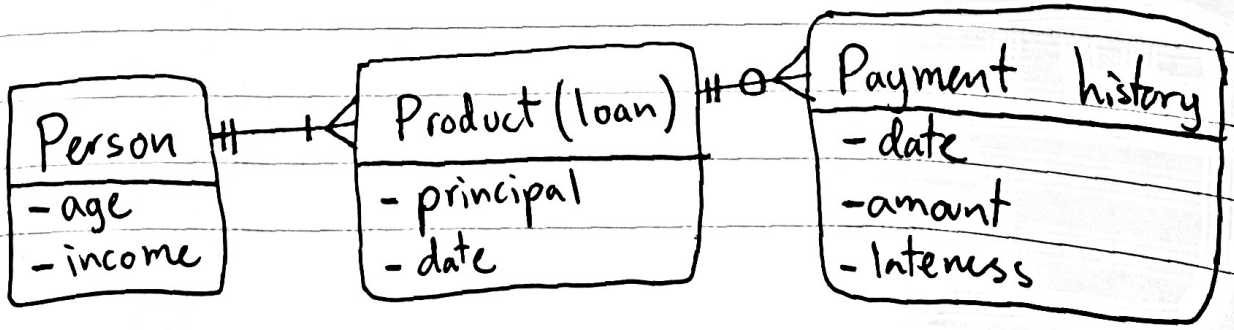


Examples

- Doctor knows $\underline{x}_n$ = (age, weight, height) and wants to learn y_n = blood pressure

- Houses with x_n = (sqm, age, floor), investor wants to know y_n = sales price

- Probability of default given products and payments (relational, tree-like structure)



Is it really machine learning?

- statistical inference ("which variables affect blood pressure")
- optimization ("find best solution / configuration" "which house is undervalued")
- agent navigating search space ("play chess")

ML = function approximation for the purpose of out-of-sample prediction

Linear regression

$$\hat{y}_n = \underline{w}^T \underline{x}_n = \sum_i w_i x_{ni} \qquad \hat{\underline{y}} = X\underline{w}$$

$$\widehat{\text{blood\_pressure}} = w_1 \cdot \text{age} + w_2 \cdot \text{weight} + w_3 \cdot \text{height}$$

$$\underline{e} = (\underline{y} - \hat{\underline{y}}) = (\underline{y} - X\underline{w})$$

$\underline{e}$ = vector with n error terms

OLS: ordinary least squares

minimize $\underline{e}^T\underline{e}$
$\underline{w}$

$$\|e\|_2^2 = \underline{e}^T\underline{e} = (\underline{y} - X\underline{w})^T(\underline{y} - X\underline{w}) = \sum_n e_n^2 = \sum_n \left[y_n - \underline{x}_n^T \underline{w}\right]^2$$

Convex optimization problem, easily solved.

All errors are not the same - norms

$$\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$$ p-norms, or ℓ-norms

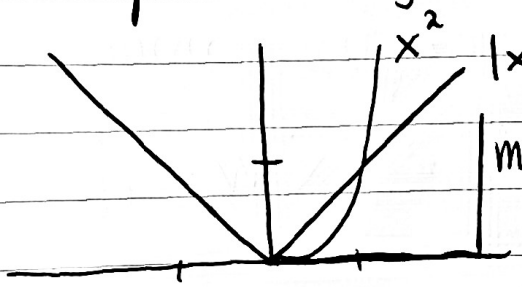
$p = 1$ Manhattan distance
$p = 2$ Euclidean distance
$p = \infty$ Max-norm

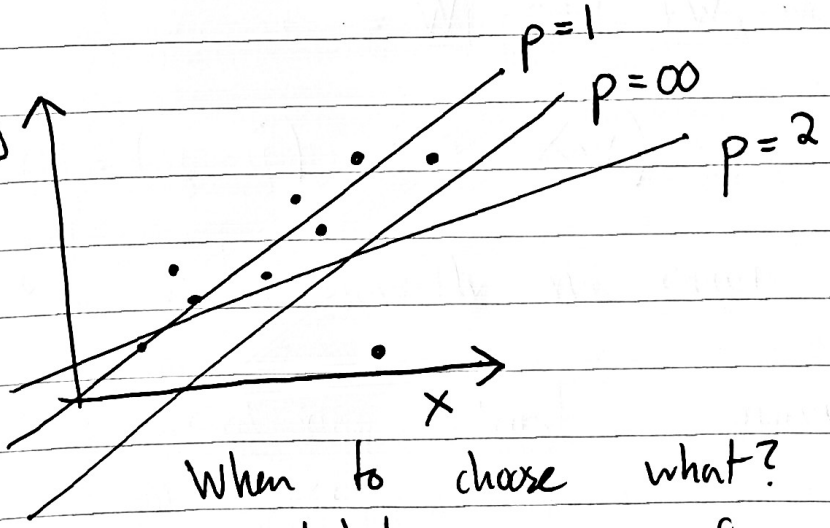
Consider

$\min_x \|x\underline{1} - \underline{a}\|$, $x \in \mathbb{R}$

median, mean, mid-range

$p = 2$ $e = (2, 2)$ better than $(4, 0)$, "fair"
$p = 1$ less sensitive to outliers
$p = \infty$ only cares about worst case (minmax)





When to choose what?

$p = 2$ distribute errors, few outliers
$p = 1$ errors tied to economics, serious outliers
$p = \infty$ make worst case as ok as possible

Feature engineering

$$\text{blood\_pressure} = \underbrace{w_1 \cdot \text{age} + w_2 \cdot \text{weight} + w_3 \cdot \text{height}}_{\text{features in data}} + \underbrace{w_4 \left(\frac{\text{weight}}{\text{height}}\right)}_{\substack{\text{new,} \\ \text{engineered} \\ \text{feature}}}$$

φ is the feature-mapping

$\hat{y}_n = \underline{w}^T \varphi(\underline{x}_n)$ instead of $\hat{y}_n = \underline{w}^T \underline{x}_n$

Example 1 Altman's financial ratios for predicting bankruptcy in 1968.

Example 2 Using "income" is a bad idea. Feature will "drift" because of inflation.

Example 3 If a variable is log-normal, taking logs often helps.

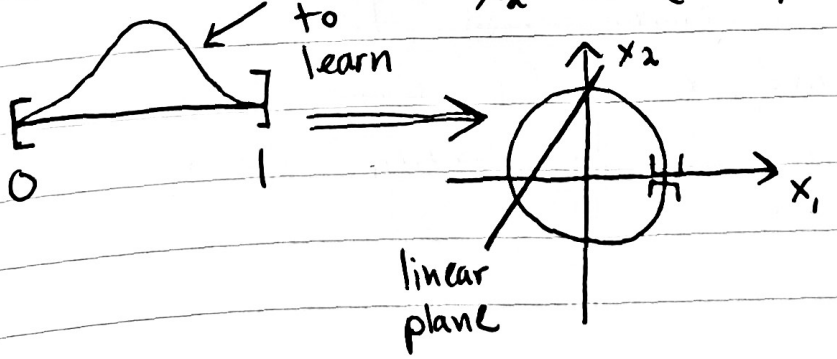
Example 4 Predicting ice-cream sales using time as a variable.

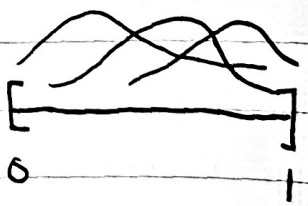
$t \in [0, 1]$ want to learn

$x_1 = \sin(2\pi t)$

$x_2 = \cos(2\pi t)$

$y = w_1 \sin(2\pi t) + w_2 \cos(2\pi t)$

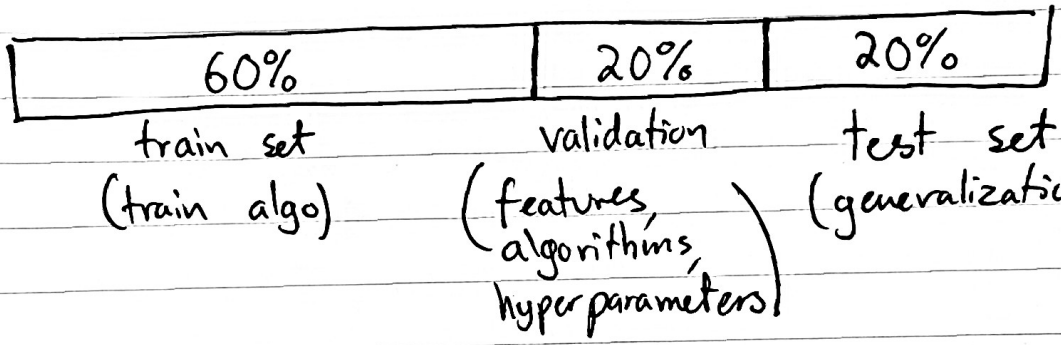




Feature engineering tips

- ask domain experts
- prefer dimensionless variables
- try logs, exp, polynomial combinations
- use relational structures, aggregate:
  - mean, variance, length, max-min, etc...

Train, validate, test sets



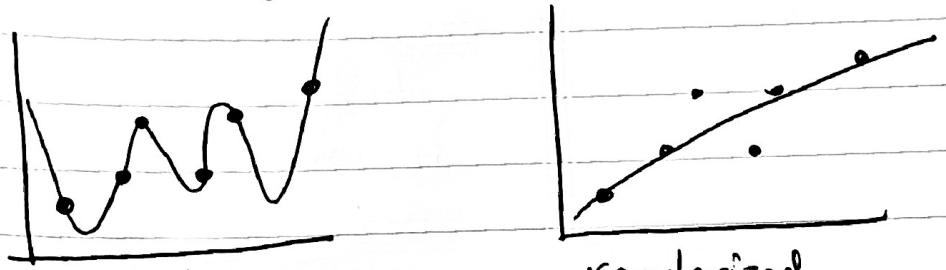
- or use k-folds cross validation

Python and ML

- pandas
- scikit-learn

Regularization

- penalize complex models to avoid overfitting (ex: little data,



overfitting regularized

$$\underset{\underline{w}}{\text{minimize}} \left(\underbrace{\|\underline{y} - X\underline{w}\|_2^2}_{\text{model fit error}}, \underbrace{\|\underline{w}\|_p^2}_{\text{model complexity}} \right)$$

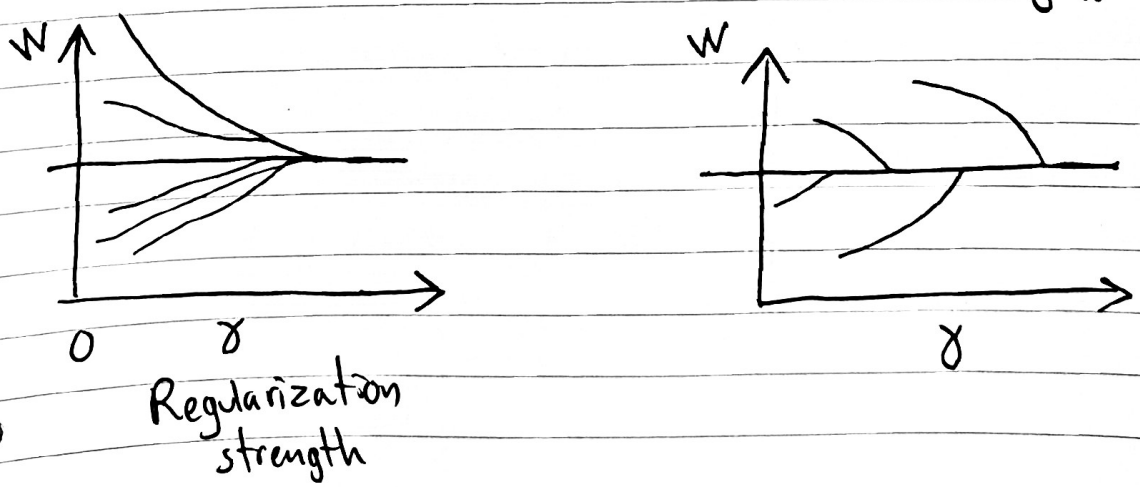
multi-objective problem, scalarized as

Ridge $p=2$

$$\min_{\underline{w}} \|\underline{y} - X\underline{w}\|_2^2 + \gamma \|\underline{w}\|_2^2$$

Lasso $p=1$

$$\min_{\underline{w}} \|\underline{y} - X\underline{w}\|_2^2 + \gamma \|w\|_1$$



Data types

- categorical ("color") [nominal] (no order)
- integer ("age") [ordinal] (rank order)
- float/real ("pressure") [interval scale]
  [ratio scale]

Examples
- grades {"A", "B", "C", ...}
- post codes {"4016", "4017", ...}

One hot encoding

color
blue
red
red
yellow

⟹

color_blue	color_red	color-yellow
1	0	0
0	1	0
0	1	0
0	0	1

- can use most common entries if there are many.

Missing data

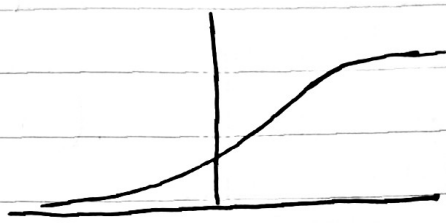
- what does missing mean? unmeasured, not existing
- is data missing at random, or does it mean anything for y? ("age" vs. "car" missing)

Strategies
- set to zero ("car age")
- set to median of column/variable
- drop rows with missing data
- KNN / multiple imputations

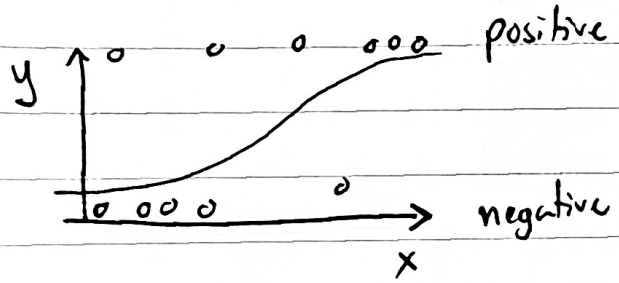
Generalized linear models

$$\hat{y}_n = f(\underline{w}^T \varphi(\underline{x}_n))$$

$$f(z) = \frac{1}{1+e^{-z}}$$



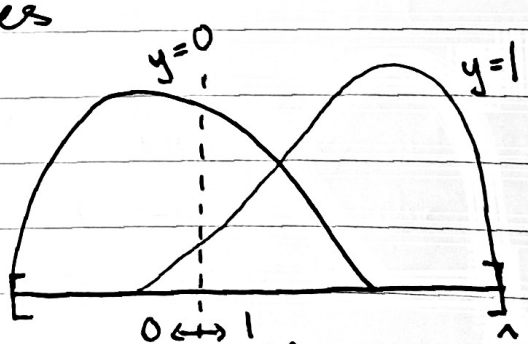
$$\min \; -\underbrace{\sum_n \ln\left[\hat{y}_n^{y_n} \cdot (1-\hat{y}_n)^{1-y_n}\right]}_{\text{log-likelihood}} + \underbrace{\|w\|_p^2}_{\text{regularization}}$$

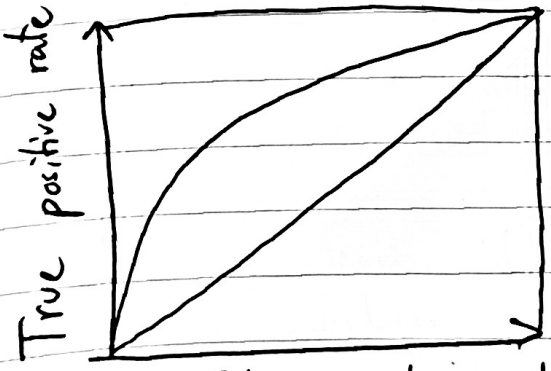


$$\hat{y} = f(w_1 + w_2 x)$$

- logistic regression is classification, but provides probabilities
- poisson

ROC CURVE (metric)





y \ ŷ	0	1
0	TN	FP
1	FN	TP

$$TPR = \frac{TP}{TP+FN} \qquad FPR = \frac{FP}{TN+FP}$$

Threshold @ 0 ⇒ all 1 ⇒ TPR = 1, FPR = 1 ⇒ top right

Sklearn pipelines

- Estimator (fit)
- Predictor (fit, predict)
- Transformer (fit, transform)

Pipelines

T → T → T → ... → T

T → T → T → ... → T → P

fit = fit, then transform, then next

transform = transform all

predict = transform all, predict

FeatureUnion can be used in parallell

Example

SimpleImputer (strategy="mean")

↓

OneHotEncoder()

↓

PolynomialFeatures (degree=2)

↓

Lasso (alpha=2)

Hyperparameters
- strategy
- degree
- alpha

Grid search
Random grid search
Bayesian Optimization

K-folds crossvalidation

Tree models

- can be used for classification / regression
- "20 questions" and splitting variables
- very interpretable ("white box" model)
- many hyperparameters

- bagging : average over many trees
- boosting : train tree n on the erros made by tree n-1.