21. Introduction to Machine Learning

金力 Li Jin
li.jin@sjtu.edu.cn
上海交通大学密西根学院
Shanghai Jiao Tong University UM Joint Institute

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Outline

- Introduction to machine learning
- Linear regression
  - EV station location
- Linear classification
  - Land use identification

Motivation for machine learning

Statistical learning plays a key role in many areas of science, finance and industry. Here are some examples of learning problems:

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.
- Predict the price of a stock in 6 months from now, based on company performance measures and economic data.
- Identify the numbers in a handwritten ZIP code, from a digitized image.
- Estimate the amount of glucose in the blood of a diabetic person, from the infrared absorption spectrum of that person's blood.
- Identify the risk factors for prostate cancer, based on clinical and demographic variables.

Typical scenario

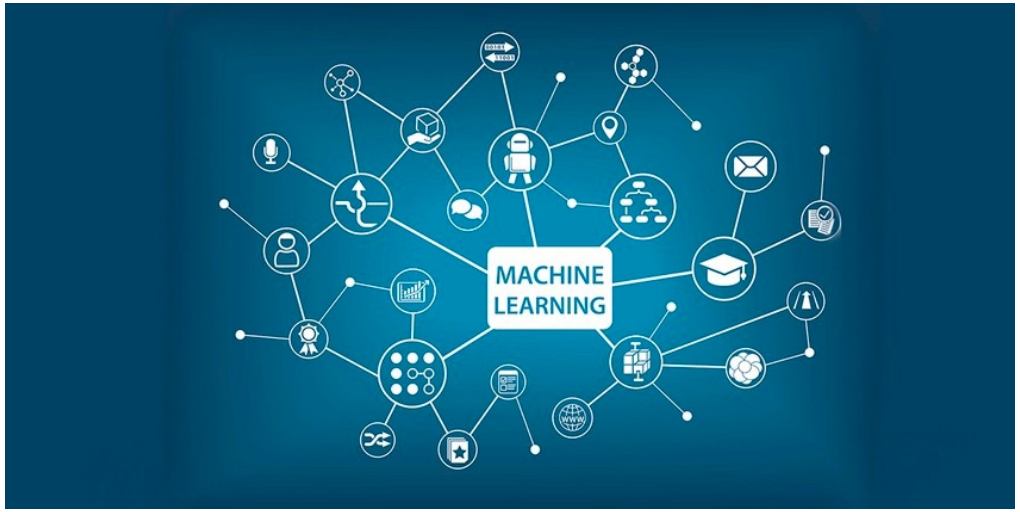
- In a typical scenario, we have an outcome measurement, usually quantitative (such as a stock price) or categorical (such as heart attack/no heart attack), that we wish to predict based on a set of features (such as diet and clinical measurements).
- We have a training set of data, in which we observe the outcome and feature measurements for a set of objects (such as people).
- Using this data we build a prediction model, or learner, which will enable us to predict the outcome for new unseen objects.
- A good learner is one that accurately predicts such an outcome.

Our focus

- The science of learning plays a key role in the fields of statistics, data mining and artificial intelligence, intersecting with areas of engineering and other disciplines.
- We describe the underlying concepts and considerations by which an engineer/researcher can judge a learning method.
- We will discuss things in an intuitive fashion, emphasizing concepts rather than mathematical details.

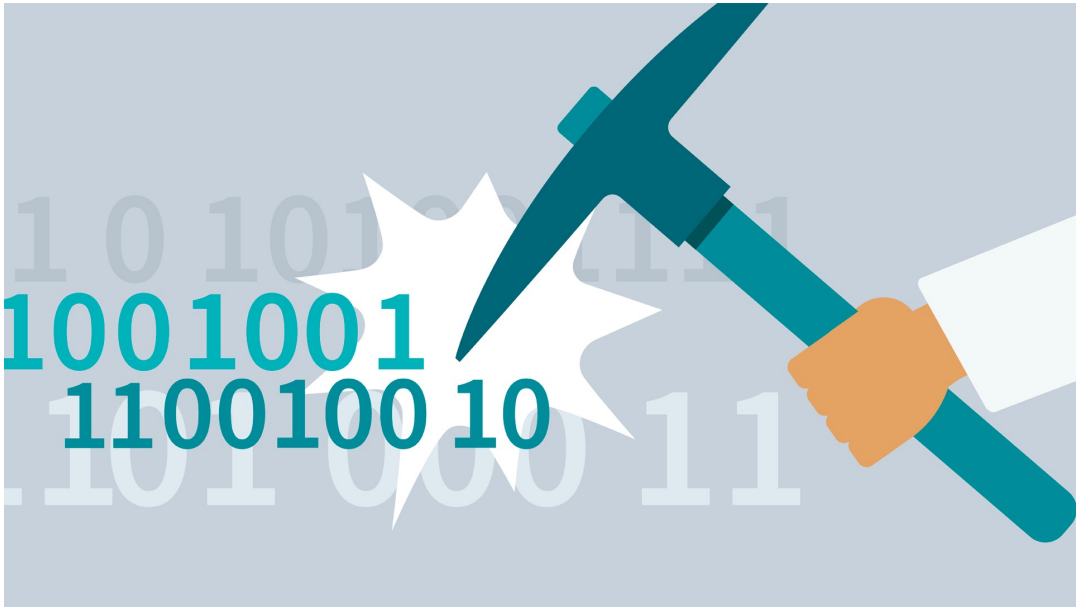
Machine learning

- Study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.
- Machine: automatic, minimal human intervention
- Learning: human do not specify everything; machine itself evolves



Data mining

- Discovery of patterns in data
- Using machine learning, statistics, and database systems
- Application of machine learning
- “Big Data”
  - Amount
  - Heterogeneity



Artificial intelligence

- Any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals
- Colloquially, the process of maximizing chance of success should be a non-trivial one
- For example, automatic transmission is responsive to environment (throttle and brake), but the response is trivial; hence this is not considered as AI
- But self-driving is responsive to the environment with a much more sophisticated mechanism, so it is AI

Machine learning algorithms

- Linear regression/classification
- Nonlinear regression/classification
- k-nearest neighbor, clustering
- Tree regression/classification
- Support vector machine
- Neural networks
- Deep neural networks

Learning process

- Initially we do not know the coefficients $\beta_0, \ldots, \beta_3$
- As time goes, we collect a growing set of data and update our knowledge of the coefficients
- We can also discount the history based how old they are: for some $\gamma \in (0,1)$

$$RSS(t) = \sum_{s=1}^{t} \gamma^{t-s} \left(\sum_{k} \left(\rho_k(s) - \hat{\rho}_k(s) \right)^2 \right)$$

- Thus, by minimizing RSS for each time step, we keep updating our knowledge of coefficients by incorporating most recent data and forgetting old data
- Evolving knowledge -> learning

Supervised learning

- The linear regression that we saw belongs to a class of learning methods called supervised learning
- We know the true response $\rho_k(t)$, and we want our predicted response $\widehat{\rho}_k(t)$ to be as close to the true one as possible (i.e. minimize RSS)
- Hence, we are "supervising" or "teaching" the training of the coefficients
- This is called supervised learning

Variable types and terminology

- Input, independent variable, predictors
- Output, dependent variable, response
- Continuous output: quantitative
  - Traffic density on a road section
  - Power consumption in a district
  - Speeds of nearby vehicles
- Discrete output: qualitative or ordered
  - Whether an email is spam
  - Whether a train passenger is a commuter or a tourist
  - Whether a vehicle is a car or a truck
  - Whether an aircraft is light, heavy, or jumbo

Training, validation, and test

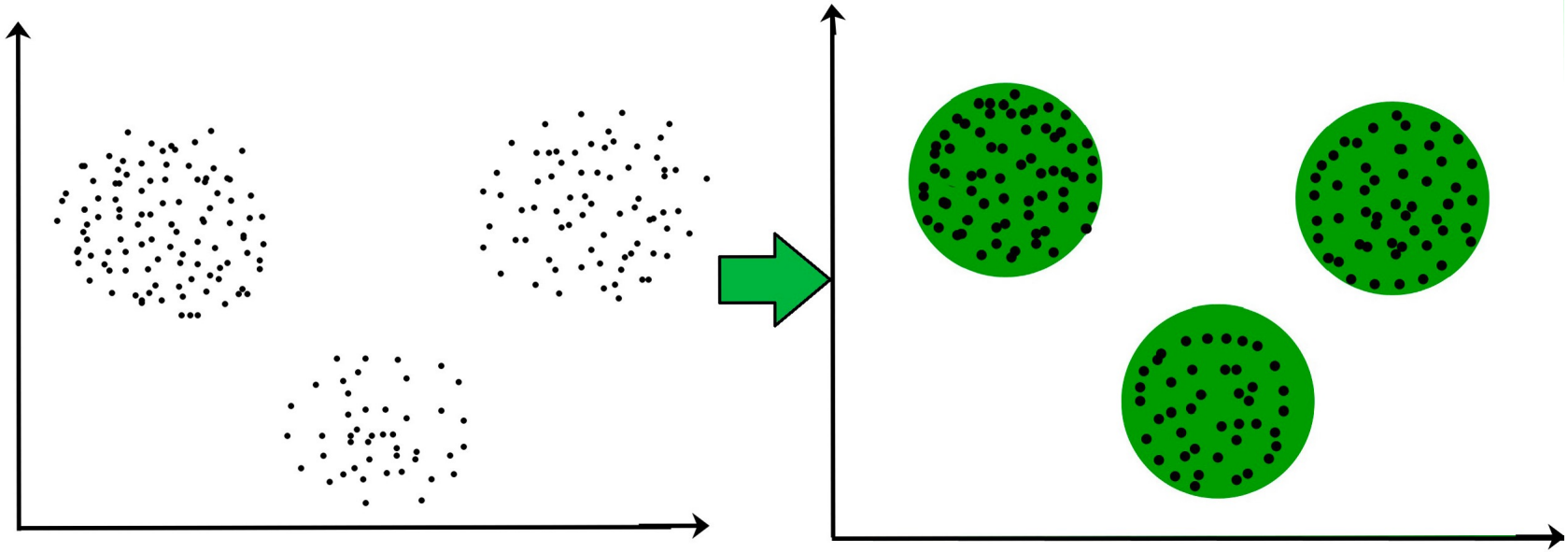
- Training: minimize RSS
- Validation: change model structure to figure out the best model
- Test: apply the validated model to a new dataset (pass or fail)
- Typically, we partition a dataset into a training set, a validation set, and a test set
- You demonstrate your test performance to the audience
- Example: linear regression

Unsupervised learning

- Supervised learning: predictions based on the training sample $(x_1, y_1), \ldots, (x_n, y_n)$ of previously solved cases
- Under this metaphor, the "student" presents an answer $\hat{y}_i$ for each x_i in the training sample
- The supervisor or "teacher" provides either the correct answer and/or an error associated with the student's answer.
- This is usually characterized by the RSS (for continuous output) or prediction error (for discrete output).
- What if we only know the input but don't know the correct answer?

Clustering

- Some points cluster and are likely to belong to a particular class
- We don't know their classes, but they appear to be, so we consider them to be.



Google PageRank

- An unsupervised method to learn a webpage's importance
- Output: ranking of a set of webpages according to their importance
- Main idea: a webpage is important if many important pages link to it
- How it works: a web crawler randomly walks over the Internet and update the importance accordingly

Generic formulation

- Predictors/Inputs/Features: $x = [x_1, x_2, \dots, x_p]^T$
- Responses/Outputs/Outcomes: $y = [y_1, y_2, \dots, y_r]^T$
- Hypothetical relation between predictors and responses:

$$y = f(x).$$

- Observed data: X, Y
- Predicted response:

$$\hat{Y} = f(X).$$

- Prediction error: difference between Y and $\hat{Y}$.

How we will discuss machine learning

- Typical algorithms their very basic forms
- Smart city-related problems to which machine learning is applicable
- You are supposed to
  - Know the mathematical definition of the above methods (excluding reinforcement learning)
  - Implement algorithms for simple examples
  - Tell the advantages/disadvantages of the above methods

Discussion

Question: is machine learning making things more interesting or less interesting?

- Banks use machine learning to detect frauds
- Doctors use machine learning to diagnose patients
- Companies use machine learning to filter job candidates
- Courts use machine learning to judge criminals
- Coaches use machine learning to select line-ups
- Movie makers use machine learning to determine cast
- Composers use machine learning to produce music of certain styles

Outline

- Introduction to machine learning
- Linear regression
  - EV station location
- Linear classification
  - Land use identification

Wagner S, Götzinger M, Neumann D. Optimal location of charging stations in smart cities: A points of interest based approach[J]. 2013.

Motivating example: electric vehicle charging

Background

- Electric vehicles (EV) have become one of the most promising transportation alternatives in recent years.
- Due to continuously increasing gas prices and CO2 taxes, while at the same time subsidies of electrified cars run into millions, many countries such as the USA, UK, and Germany intend to bring large amounts of EVs onto their roads soon.
- As a prerequisite, an adequate charging infrastructure is needed to supply these vehicles with electrical fuel.
- However, planning EV charging stations is complex and not straightforward.

Where to locate the charging stations?

Key: demand prediction

- What is a good location for a charging station?
  - Supermarket?
  - Shopping mall?
  - Hospital?
  - Bank?
  - Train station?
  - Museum?
  - School?
- So many competing factors; how to compare?
- An even harder question: how to quantify?
- Linear regression!

Some terminologies

- Independent variable or predictor or feature x
  - Something that we consider to be the "cause" or "determining factor"
  - Something that we know
- Dependent variable or response y
  - Something that we consider to result from the predictor
  - Something that we want to predict

Hypothetical linear relation

- Linear regression hypothesizes that response variable linearly varies with independent variable, i.e.

$$y = \beta_0 + \beta_1 x + e$$

- β_0 is the intercept, β_1 is the coefficient, e is a zero-mean random variable called the noise
- Noise = randomness due to our lack of information
- We assume that e follows normal distribution
- This is a hypothetical relation, which may or may not be true
- For EV charging demand prediction
  - x = land use, population density, neighborhood
  - y = demand for EV charging

When is linear relation a good assumption?

- Response variable is monotonic in independent variable
- Charging demand is monotonic in
  - Land use? (binary indicator variable)
  - Population density?
  - Density of public transit service?
  - Annual average temperature?
- Look at data and see what data say

Fitting a straight line

- Suppose that n repeated experiments have been done
- The experiments use $x_1, x_2, \dots, x_n$ as the values of the dependent variables.
- The experiments lead to the following results: $y_1, y_2, \dots, y_n$.
- Hypothesize that $y = \beta_0 + \beta_1 x + e$
- What should β_0 and β_1 be? -> least squares

Method of least squares

- Given β_0 and β_1, the prediction error of the linear model $y = \beta_0 + \beta_1 x$ is

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- Principle of method of least squares: find parameters β_0 and β_1 that minimize the sum of prediction errors

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min S(\beta_0, \beta_1)$$

- Least square!
- We will discuss the math next time

Multivariate linear regression

- Consider $p-1$ independent variables $x_1, x_2, \ldots, x_{p-1}$
- One dependent variable y
- Hypothetical relation

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{p-1} x_{p-1} + e$$

- To determine the intercept and coefficients, minimize the mean square error

$$S = \sum_{i=1}^{n} (y_i - \quad (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_{p-1} x_{i,p-1}) \quad)\text{^}2$$

Multivariate linear regression

- Consider $p-1$ independent variables $x_1, x_2, \ldots, x_{p-1}$
- One dependent variable y
- In matrix form

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,\,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,\,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,\,p-1} \end{bmatrix}$$

- Be careful: X begins with a column of 1's!
- Hypothetical relation

$$Y = X\beta$$

- So we want to minimize MSE $S(\beta) = \|Y - X\beta\|^2$
- $\|v\|^2 = v_1^2 + v_2^2 + \cdots + v_m^2$

Linear least squares

- In order to fit a straight line to a plot of points (x_i, y_i), where $i = 1, \ldots, n$, the slope and intercept of the line $y = \beta_0 + \beta_1 x$ must be found from the data in some manner.
- In order to fit a pth-order polynomial, $p + 1$ coefficients must be determined.
- Other functional forms besides linear and polynomial ones may be fit to data, and in order to do so parameters associated with those forms must be determined.

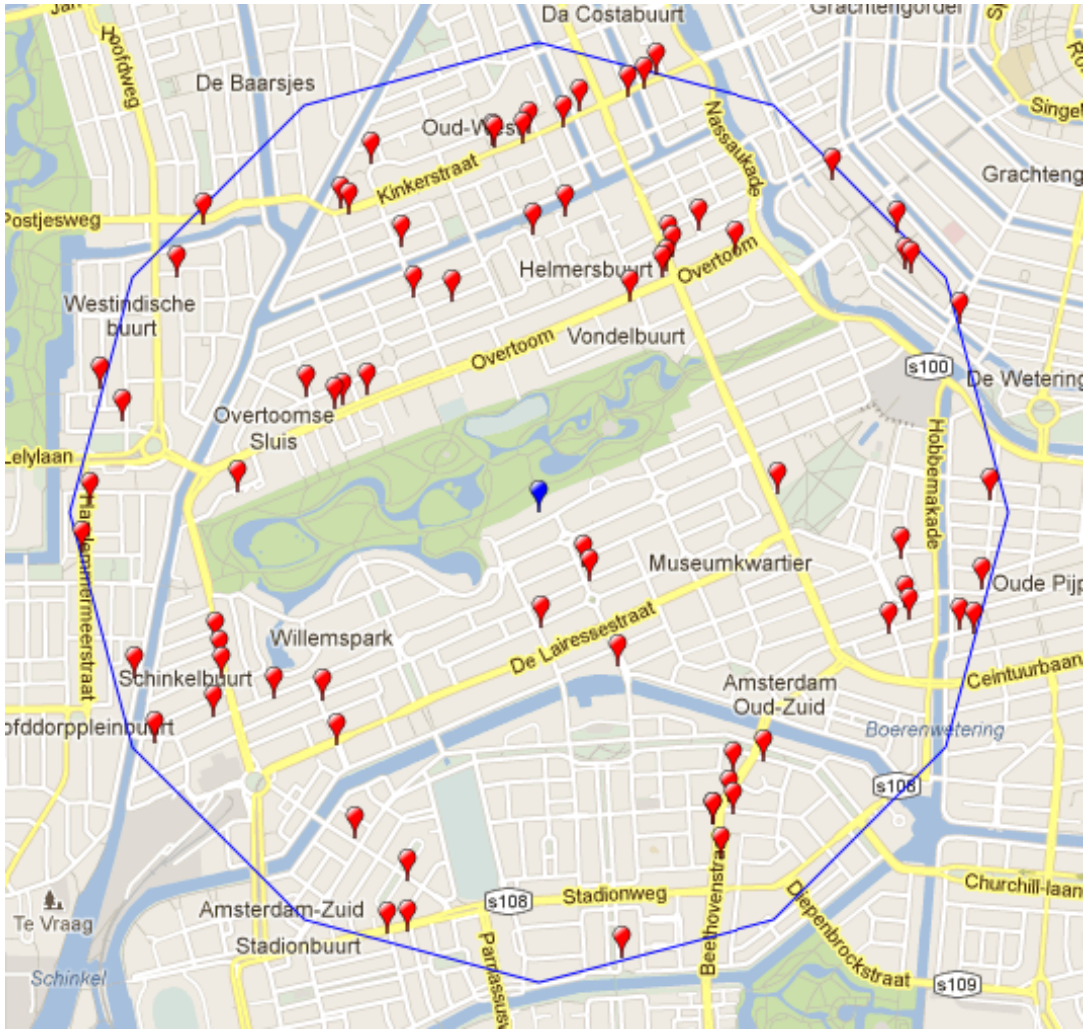
Formulae for model parameters

- The formula of least square estimation:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- To implement this, you may want to refresh your knowledge about
  - Matrix multiplication
  - Transpose
  - Inverse
  - How to implement them in Excel/Python/MATLAB/c++…
  - For example, in Excel we have mmult and minverse
  - Python/MATLAB are simpler

Application: EV charging station allocation



Background

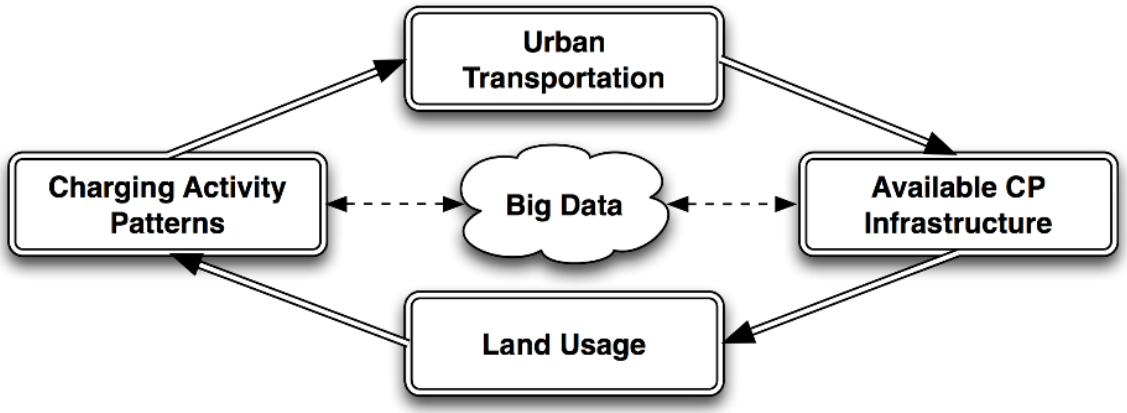
- Electric vehicles (EVs)
- Greener than fueled vehicles
- Enabled by developing battery technology
- A major challenge: range anxiety
  - Am I running out of power?
  - Where is the nearest charging station?
  - Is my destination too far away from a charging station?

Planning charging stations

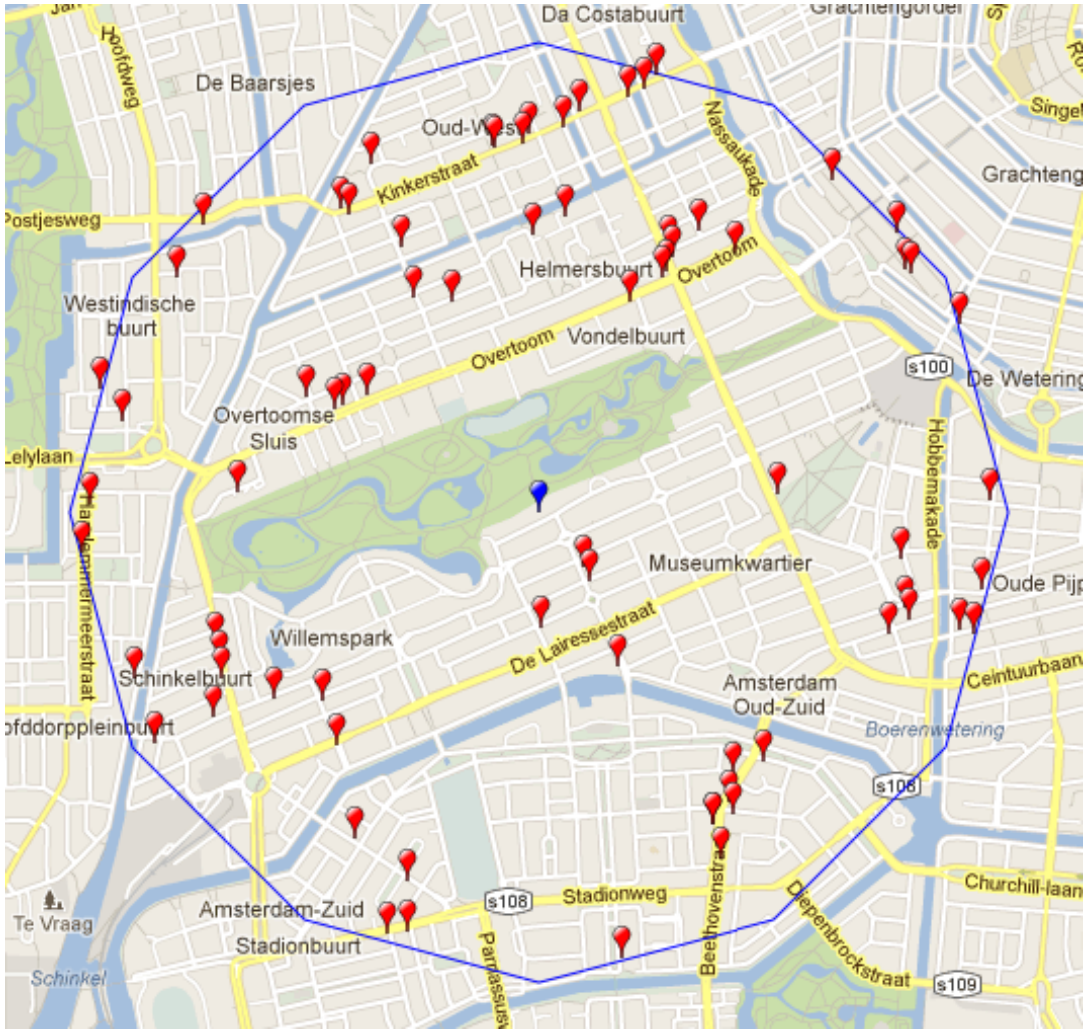
- Support city planners to locate charging stations
- Planning based on big data
- Other buzzwords involved:
  - Business intelligence
  - e-government
  - urban economics

The planning strategy

- User-centric
  - accounting for points of interest
- Data-driven
  - done by linear regression
- Basic principle: locate charging stations at popular points of interest



Fitting LR model for data

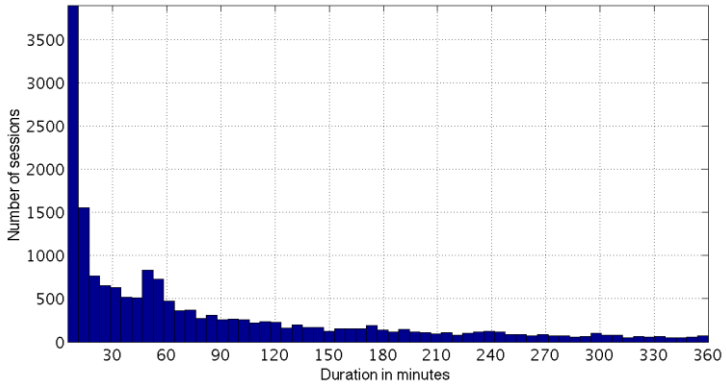


Dependent variable (response)

- Charging station importance y

$$y = \frac{1}{n}\sum_{d=1}^{n} \#\ of\ uses \times average\ duration\ on\ day\ d$$

- Basically the daily average number of vehicle-minutes
- The higher the usage, the more important the charging station will be
- Observations collected from historical data



Independent variable (predictor)

- Position of a charging station (latitude and longitude)
- Type: binary indicating variables
  - Supermarket?
  - Shopping mall?
  - Hospital?
  - Bank?
  - Train station?
  - Museum?
  - School?
- Distance to closest charging station (km)
- All the above quantities make up the vector of independent variables x

Linear regression

- Hypothetical relation: $y = \beta^T x + e$

Table 1. POI Regression

Regression Statistics	
R^2	0.164
Adjusted R^2	0.101
F-value	2.593
p-value	0.001

	Coefficients	t-statistic	Significance[1]	p-value
Intercept	0.31	5.72	***	0.000
Food	0.55	2.14	**	0.034
Store	-0.66	-2.21	**	0.028
Health	0.56	2.55	**	0.012
Finance	-0.18	-1.47		0.144
Bus station	-0.12	-1.68	*	0.095
Museum	0.35	2.15	**	0.033
School	-0.30	-1.92	*	0.056
Church	-0.02	-0.16		0.874
Travel agency	-0.03	-0.19		0.850
Hair care	-0.03	-0.26		0.794

1. significance: *** 99%, ** 95%, 90%, blank <90% Observations: 229

Observations

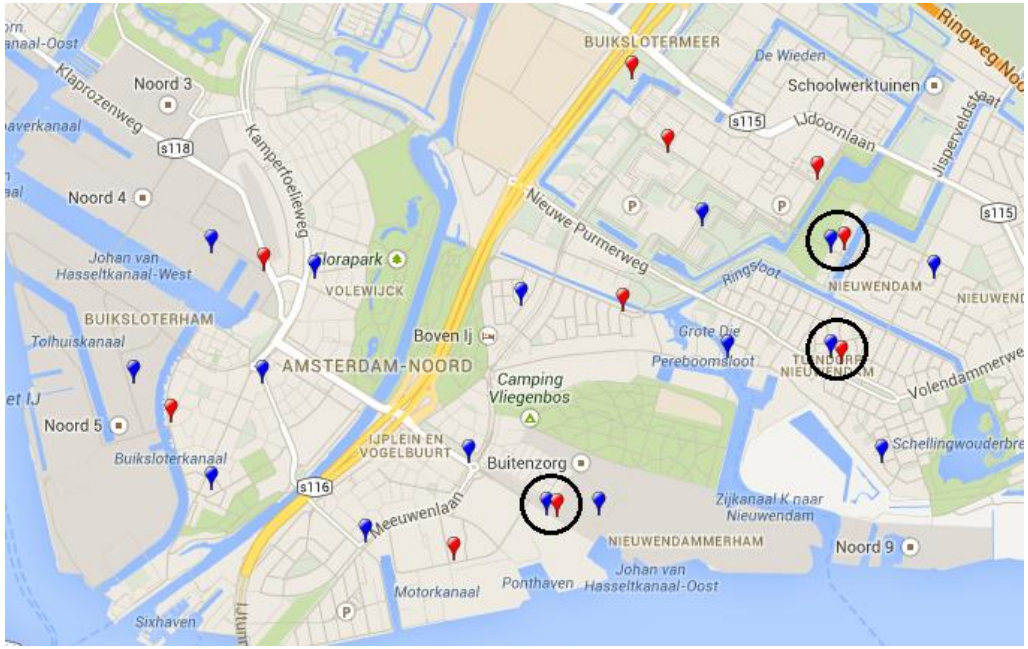
- Category types *food*, *health*, and *museum* show a significance value of more than 95%, with a t-statistic value greater than 2, which indicates a positive influence on charge point usage.
- Since, visiting a museum is a time consuming activity, the charging session durations of EVs will be substantially higher in contrast to withdrawing money from an ATM.
- This is also the reason, why e.g. the *finance* category type provides no significance at all.
- Overall, POIs have a significant influence on the usage of charge points and, thus, have to be considered when developing a future charging infrastructure for smart cities.

Planning stations using the LR model

- decision variable:
  - locations of p charging stations
- maximize:
  - covered demand
- subject to:
  - coverage radius of a station
  - geographical constraints on station locations

Case study for Amsterdam

- Red = current locations
- Blue = proposed locations



Test-retest

- The regression effect must be considered in test-retest situations.
- Suppose, for example, a location is selected for EV charging station
- Then, people will adapt themselves to the EV station
  - More trips made to this location
  - Land use may be altered
- However, such increase results from the existing EV station, not from independent variables.

Outline

- Introduction to machine learning
- Linear regression
  - EV station location
- Linear classification
  - Land use identification

Zhu Y, Newsam S. Land use classification using convolutional neural networks applied to ground-level images[C]//Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2015: 1-4.

Classification

- Predictors: $x = (x_1, x_2, \dots, x_p)$
- k features of a sample
- Response: $G \in \{1,2, \dots, K\}$
- Every sample belongs to one out of K classes
- Classification problem: given predictors $X = x$, what is the probability that this sample belongs to a particular class?

$$\Pr\{G = k | X = x\} = ?$$

- If we can compute this probability, then our prediction of class is

$$\hat{G} = \arg\max_k \Pr\{G = k | X = x\}$$

Example

Electronic toll collection (ETC):

- Predictors: weight, length, speed
- Response: car? Bus? Truck?

Building access control:

- Predictors: image (RGB pixels)
- Response: resident? Staff? Unauthorized person?

Linear regression of an indicator matrix

- Predictors: $x = (x_1, x_2, \dots, x_p)$
- Response: $y = (y_1, y_2, \dots, y_K)$
  - If an observation is in class k, then $y_k = 1$ and $y_j = 0$ for $j \neq k$
- Assume that the indicator variables $y_1, y_2, \dots, y_K$ linearly depend on the predictors
  - $y_1 = \beta_{1,0} + \beta_{1,1}x_1 + \beta_{1,2}x_2 + \cdots + \beta_{1,p}x_p$
  - $y_2 = \beta_{2,0} + \beta_{2,1}x_1 + \beta_{2,2}x_2 + \cdots + \beta_{2,p}x_p$
  - …
  - $y_K = \beta_{K,0} + \beta_{K,1}x_1 + \beta_{K,2}x_2 + \cdots + \beta_{K,p}x_p$
- In matrix form
  - $y = X^T B$
  - y is $K \times 1$, B is $(p+1) \times K$, X is $(p+1) \times 1$

Linear regression of an indicator matrix

- How to determine coefficient matrix B?
- Suppose that we have n observations
- RSS = $\sum_{i=1}^{n}\sum_{k=1}^{K}\left(y_{i,k}-\left(\beta_{k,0}+\beta_{k,1}x_1+\beta_{k,2}x_2+\cdots+\beta_{k,p}x_p\right)\right)^2 = (Y-X^TB)^T(Y-X^TB)$
- Minimize RSS

$$\hat{B} = (X^TX)^{-1}X^TY$$

- Fitted classes

$$\hat{Y} = X^T\hat{B} = X(X^TX)^{-1}X^TY$$

Prediction

A new observation with input x is classified as follows

1. Compute the fitted output $\hat{f}(x) = (1, x^T)\hat{B}$, a K vector
2. Identify the largest component and classify accordingly

$$\hat{G}(x) = \arg\max_k \hat{f}_k(x)$$

Why?

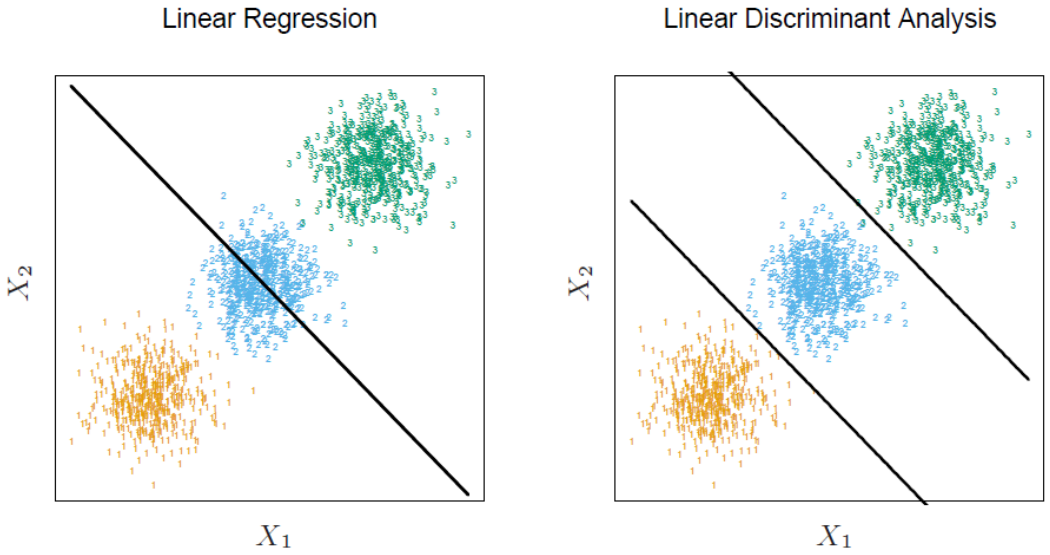
- A rather formal justification: y_k is a random variable such that

$$E[\hat{y}_k | X = x] = \Pr\{G = k | X = x\}$$

- That is, $\hat{y}_k \approx \Pr\{G = k | X = x\}$
- Illustration for two-class problem

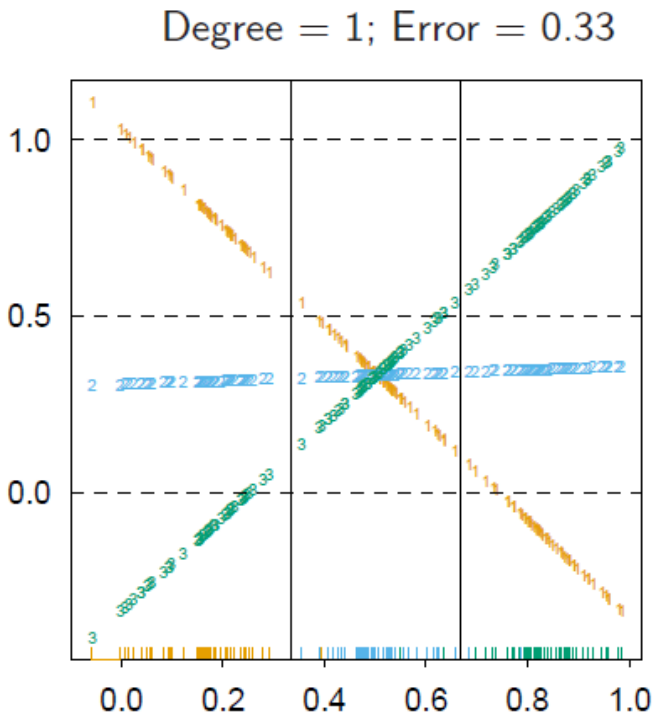
LR not always work

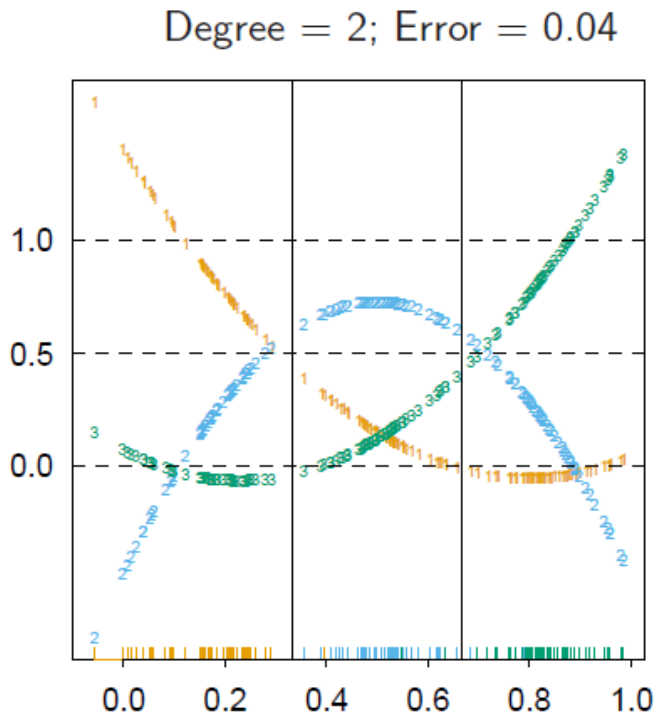
- It seems that $\hat{y}_k \approx \Pr\{G = k | X = x\}$
- However, $\hat{y}_k = \beta_{k,0} + \beta_{k,1}x_1 + \beta_{k,2}x_2 + \cdots + \beta_{k,p}x_p$ can be < 0 or > 1, which is not allowed for probabilities
- Another serious problem: masking
- Illustration for three-class problem



Solution: non-linear methods

- Linear discriminant analysis (LDA)
- Logistic regression
- Only need to know what they mean





Linear discriminant analysis

- Consider N observations
- Predictor vector x
- Class g
- Define
  - $\hat{\pi}_k = N_k/N$, where N_k is the number of class-k observations
  - $\hat{\mu}_k = \sum_{g_i=k} x_i/N_k$
  - $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{i:g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_i)^T/(N-K)$
- Linear discriminant function

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log \pi_k$$

- Classification

$$G(x) = \arg\max_k \delta_k(x)$$

Logistic regression

- Logit function

$$\Pr\{G = k | X = x\} = \frac{\exp\left(\beta_{k0} + \beta_k^T x\right)}{\sum_{l=1}^{K} \exp(\beta_{l0} + \beta_l^T x)}$$

- Classification

$$G(x) = \arg\max_k \frac{\exp\left(\beta_{k0} + \beta_k^T x\right)}{\sum_{l=1}^{K} \exp(\beta_{l0} + \beta_l^T x)}$$

- Process of fitting coefficients β_{ki}: called logistic regression
- Use maximum likelihood

Summary of linear classification

- Predictor x
- Find some function $f_k(x)$ such that
$$f_k(x) \approx \Pr\{G = k | X = x\}$$
- For a new observation with input x, classify as follows
$$\hat{G}(x) = \arg\max_k f_k(x)$$
- $f_k(x)$ is linear in x -> linear regression
- $f_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$ -> LDA
- $f_k(x) = \frac{\exp(\beta_{k0} + \beta_k^T x)}{\sum_{l=1}^K \exp(\beta_{l0} + \beta_l^T x)}$ -> logistic regression

MATLAB coding

```
clear
close all
clc

X=[1 2 3 4 5 6]';
Y=[1 1 1 0 0 0]';

Mdl =
fitclinear(X,Y,'Learner','logistic');

X2=1.5;
Label = predict(Mdl,X2);
```

Example: Land Use Classification

Example: Land Use Classification

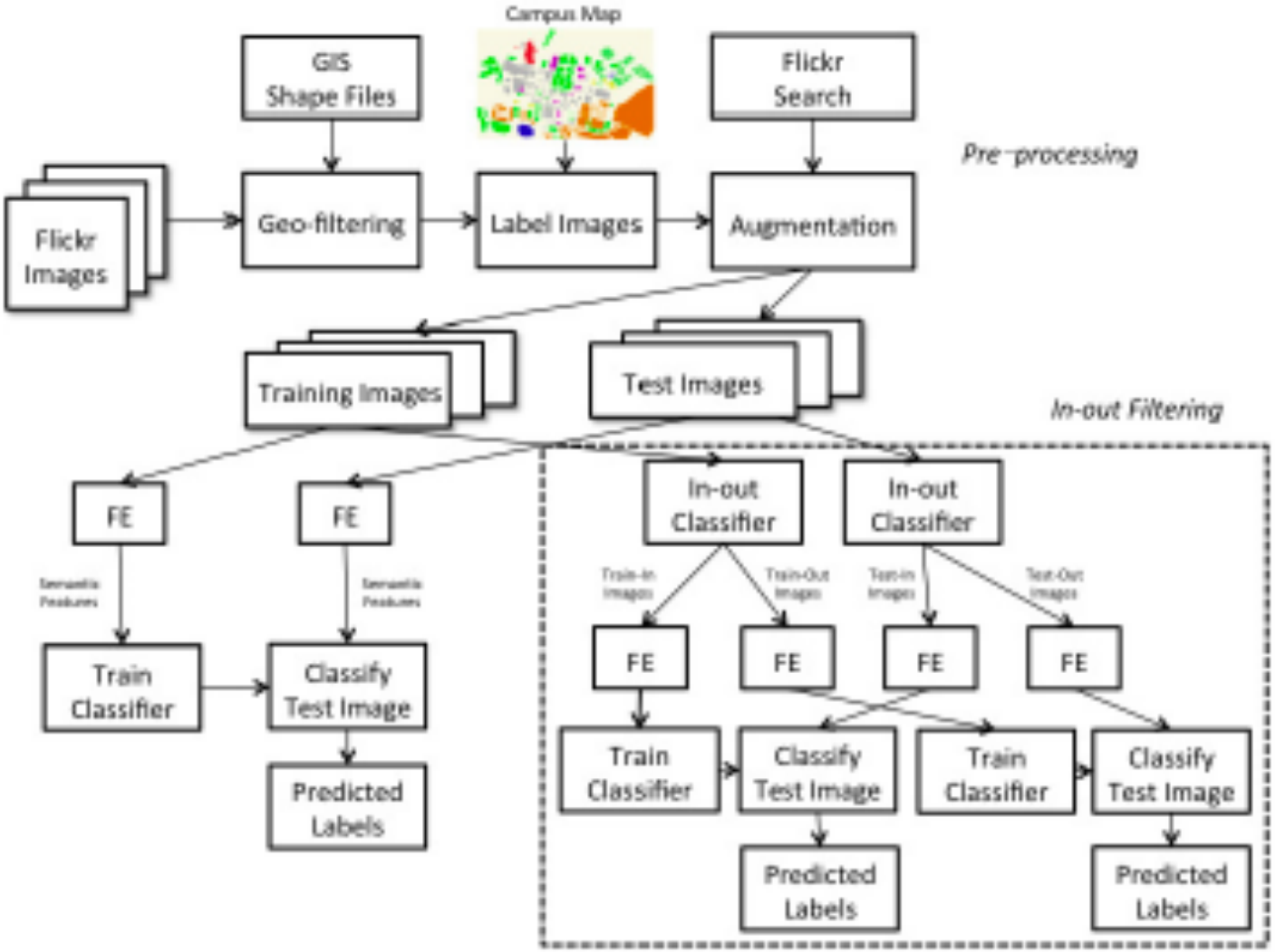
Focus: challenging problem of mapping land use.

- Map a broader range of land use classes than previous work
- Utilize semantic image features learned by training convolutional neural networks, a form of deep learning on a large collection of scene images.
- Develop an indoor/outdoor image classier which achieves state-of-the-art performance. It helps correct for image location errors.
- Region shape are used to further correct for image location errors as well as to create precise maps.
- A base set of training images is generated in an automated fashion and then augmented in a semi-supervised fashion to address class imbalance.

Overview

- Focus on land use classification on a university campus (Stanford) since it represents a compact region containing a range of classes
- manually generating a ground truth is feasible.
- Study, Residence, Hospital, Park, Gym, Playground, Water and Theater.

Workflow


Campus Map
GIS Shape Files
Flickr Search
Pre-processing
Flickr Images
Geo-filtering
Label Images
Augmentation
Training Images
Test Images
In-out Filtering
FE
FE
In-out Classifier
In-out Classifier
Train Classifier
Classify Test Image
Predicted Labels
FE
FE
FE
FE
Train Classifier
Classify Test Image
Train Classifier
Classify Test Image
Predicted Labels
Predicted Labels

Data

- use the Flickr API to download images located within the campus region
- each downloaded image is assigned a land use label according to its geographic location on the ground truth map.
- 16,789 images
- Augmentation:
  - Original data is unbalanced in terms of land use and location
  - Such imbalance can lead to a biased classifier (very critical issue!)
  - To address this, artificially adjust the ratio
  - Reduce the more frequent ones, increase the less frequent ones

Land use classification

- Use a $f_k(x)$ model called support vector machine (SVM)
- Basic idea: separate point groups that are far away
- 65% prediction accuracy
- Fine-tune SVM with validation set

Further refinement

- Two layers of classification (hierarchical)
- First layer: indoor or outdoor
  - Indoor: study, residence, hospital, gym, theater
  - Outdoor: park, playground, water
- Second layer: specific land use
- Prediction accuracy: 99%, 76%

Summary

- Introduction to machine learning
- Linear regression
  - EV station location
- Linear classification
  - Land use identification