Machine Learning for Large Scale Farming

Zachariah Litzinger

UMN Morris

April 15, 2017

What are we doing?

- Large space
- Few experts
- Sorting into types is challenging
- Need identification is challenging
- Prevent pre-treating

https://flic.kr/p/8fH39P

Outline

1. Introduction
   - Phenotyping
   - Detecting Needs
   - Collecting Data
2. Supervised Learning: Support Vector Machines
3. Unsupervised Learning: k-Means Clustering
4. Results
5. Conclusion

Phenotyping

- Sorting by type
- Many uses
  - Weeds
  - Sort species
- Example: Sorting organic and conventionally grown wheat [KBA+15]

https://flic.kr/p/6Qa14W

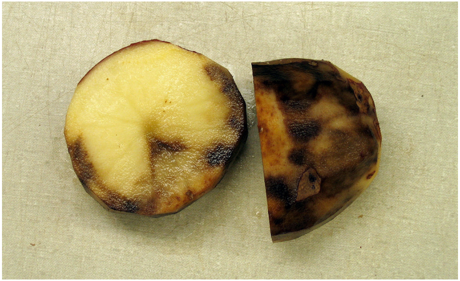
Detecting Needs: What Plants Need

- Nutrients
  - Water
  - Soil Quality
- Disease treatment

https://flic.kr/p/dBdzj

Detecting Needs

- Early recognition
- Hard to do by hand
- Increase efficiency
- Can discover new needs or detect previously known issues
- Prevent pre-treating
- Example: Detect Blight Disease in potatoes [PGHG15]
  - Irish Potato Famine
  - Still prevalent today



https://flic.kr/p/qRXVt

Collecting Data: Before

- Sampling
- Lots of person-power
- Takes a lot of time

https://flic.kr/p/81ocW2

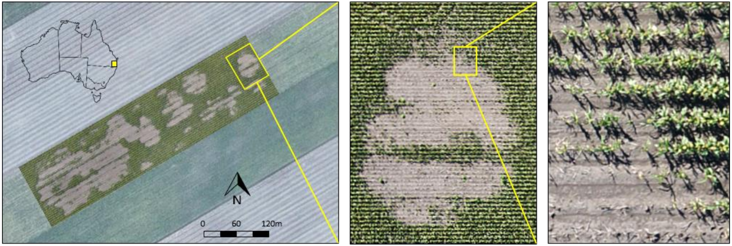
Collecting Data: Future

- Less person-power
- Faster
- Covers a lot of area, 48 minutes for 70 acre area [PGHG15]

https://flic.kr/p/ExQeNH

Collecting Data: Processing

- Reduce resolution to improve computational cost [PGHG15]
- Single picture: RGB per pixel
- Multiple pictures: scaled RGB summary per image (average)



[PGHG15]

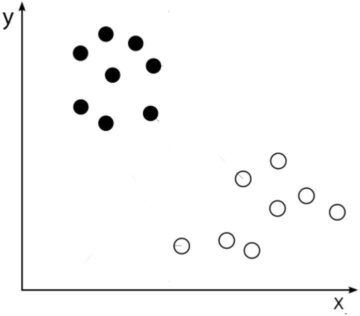
- Less data needed than traditional data analysis methods [BMR+15]
- Note: Kessler et al. analyzed already harvested corn, used metabolic data for classification

Supervised Learning

- Used to detect specific classes of data
- Requires predefined categories
- Methods used for this field:
  - Support Vector Machines
  - Random Forests
  - Neural Networks

Supervised Learning

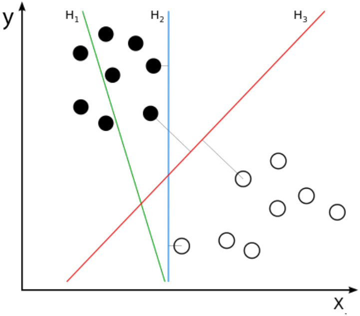
- Classify our initial data
- Split the data into two parts:
  - training data
  - testing data
- Train the method on training data
- Measure success by percent correct for testing data
- If accurate enough, we can classify future data



https://z.umn.edu/svm-linear-classifier

Support Vector Machines (SVMs)

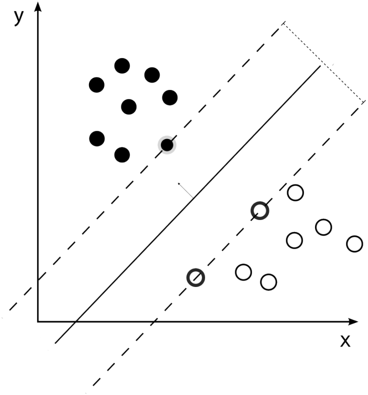
- Data is separable by a vector, so it is linearly separable by a linear classifier
- Can be high dimensional classifier (hyperplane classifier)
- H_1 is bad
- H_2 isn't optimal
- H_3 is what we are looking for



https://z.umn.edu/svm-linear-classifier

Method: SVM - Getting our Linear Classifier

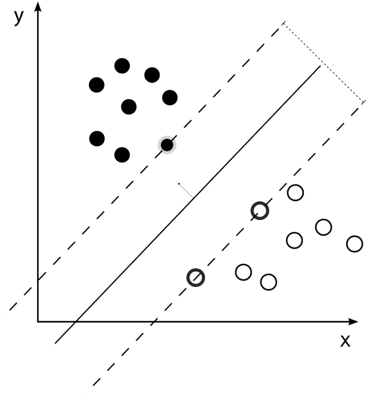
- Choose two parallel lines that separate the data
- Make them as far apart as possible
- These support vectors can be of standard line form for our purposes $ax + b = y$ where both support vectors have the same slope



https://z.umn.edu/svm-support-vectors

Method: SVM - Getting our Linear Classifier

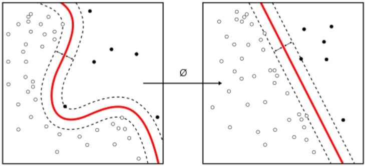
- The bisecting vector is our defining vector, also of standard line form.
- Note: In higher dimensions we use vectors for these definitions.



https://z.umn.edu/svm-support-vectors

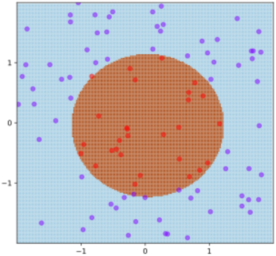
Method: SVM - Kernel Functions

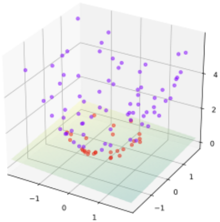
- Takes original position values
- Gives new position
- General Form:
  $\Phi(x, y) = \langle \phi(x), \phi(y) \rangle$ [mW18]
- Challenging to identify and come up with



https://z.umn.edu/kernel-machine

Method: SVM - Kernel Example





https://z.umn.edu/kernel-trick-quadratic

- Quadratic Kernel
- $\Phi(x, y) = \langle x, y, x^2 + y^2 \rangle$
- Example: $\Phi(3, 2) = \langle 3, 2, 3^2 + 2^2 \rangle = \langle 3, 2, 13 \rangle$

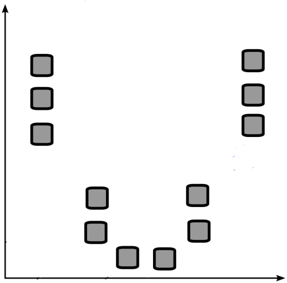
[kW18]

Unsupervised Learning

- Used to create new classes for data
- Requires analyzing new categories
- Methods used for this field:
  - k-Means Clustering
  - Image Segmentation

Unsupervised Learning

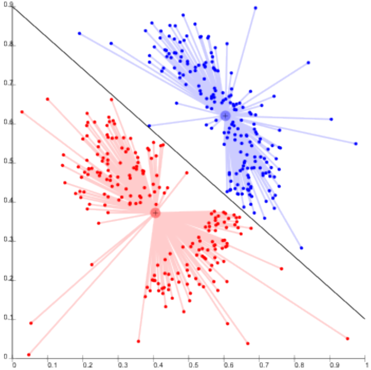
- No need to label the data
- Method classifies our data
- We analyze each of the classifications to understand what the significant property found is.



https://z.umn.edu/k-means-cluster-examples

k-Means Clustering

- Looking to cluster data into regions of most similar points
- User chooses how many regions

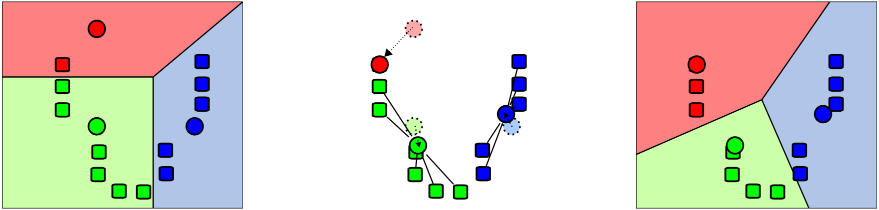


https://z.umn.edu/k-means-cluster

k-Means Clustering

- Randomly choose k points, 3 in example
- Regions defined by Euclidean Distance
- Find centroid of points in each region
- Reclassify points for new regions
- Repeat until centroid is stable [kmcW18]

$$C = \frac{\sum_{n=1}^{p} x_n}{p}$$



https://z.umn.edu/k-means-cluster-examples

Results: SVM - Classifying Corn

- Kessler et al. wanted to classify conventionally grown and organically grown wheat
- Uses 313 total samples
- Used metabolic profile instead of image data
- Goal: Create a reliable classification method for sorting when bio-markers of data are unknown, making classical statistical analysis impossible

Results: SVM

Kessler et al. Results [KBA+15]		
Year Trained On	Year Tested On	Accuracy
2007	2007	0.9677
2010	2010	0.8846
2007	2010	0.5547
2010	2007	0.5562
2007, 2009, 2010	2007, 2009, 2010	0.9032

- Shows the accuracy of same years is above or close to .9
- Cross-year results accuracy only around .55
- Outperformed statistical analysis of full bio-marker set
- Note: Behmann et al. identified that using RGB image data would be a sufficient set of input data for goals similar this. [BMR+15]

Results: k-Means

- Puig et al. wanted to detect insect damage
- Covered 70 acres of land
- Used overhead image data
- Goal: Create a “near real-time assessment” of problem spots in sorghum fields.

Results: k-Means

- Using a k-Mean value of k=3
- Successfully identified
  - Dead portions
  - Unhealthy portions
  - Healthy portions

[PGHG15]

Conclusion

With Machine Learning:

- Cover large area
- Need fewer experts
- Accurately identify needs
- Sort plants based on type
- Increase efficiency of farms

https://flic.kr/p/8fH39P

Questions?

Nic McPhee https://flic.kr/p/5aSKLx

Acknowledgments

Special thanks to Nic McPhee and Elena Machkasova for their guidance and support during this research paper.

Jan Behmann, Anne-Katrin Mahlein, Till Rumpf, Christoph Römer, and Lutz Plümer, *A review of advanced machine learning methods for the detection of biotic stress in precision crop protection*, Precision Agriculture **16** (2015), no. 3, 239–260.

Nikolas Kessler, Anja Bonte, Stefan P. Albaum, Paul Mäder, Monika Messmer, Alexander Goesmann, Karsten Niehaus, Georg Langenkämper, and Tim W. Nattkemper, *Learning to classify organic and conventional wheat – a machine learning driven approach using the meltdb 2.0 metabolomics analysis platform*, Frontiers in Bioengineering and Biotechnology **3** (2015), 35.

k-means clustering Wikipedia, *k-means clustering — Wikipedia, The Free Encyclopedia*, 2018, [Online; accessed 22-February-2018].

Polynomial kernel Wikipedia, *Polynomial kernel — Wikipedia, The Free Encyclopedia*, 2018, [Online; accessed 22-February-2018].

Kernel methods Wikipedia, *Kernel methods for vector output — Wikipedia, The Free Encyclopedia*, 2018, [Online; accessed 22-February-2018].

Eduard Puig, Felipe Gonzalez, Grant Hamilton, and Paul Grundy, *Assessment of crop insect damage using unmanned aerial systems: A machine learning approach*, 21st International Congress on Modelling and Simulation (MODSIM2015) (Gold Coast, Qld), December 2015.