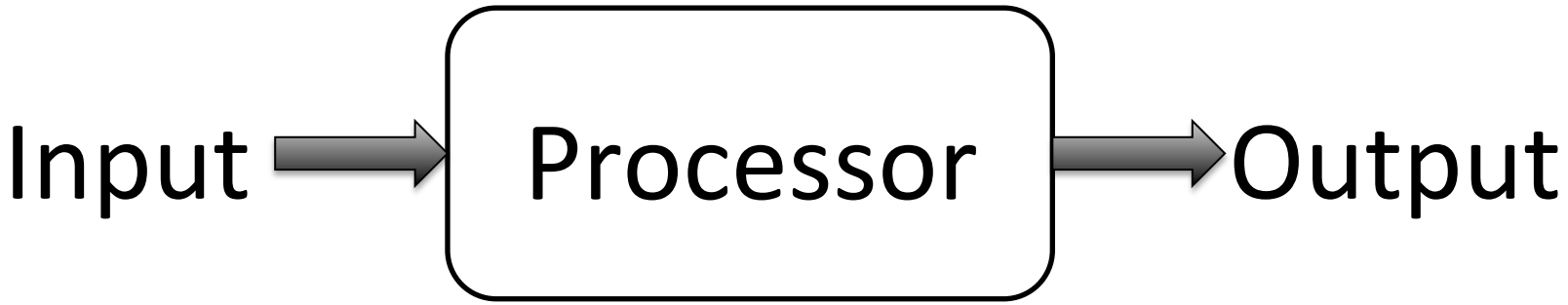
Week 3
Machine Learning Refresher

Data Analysis

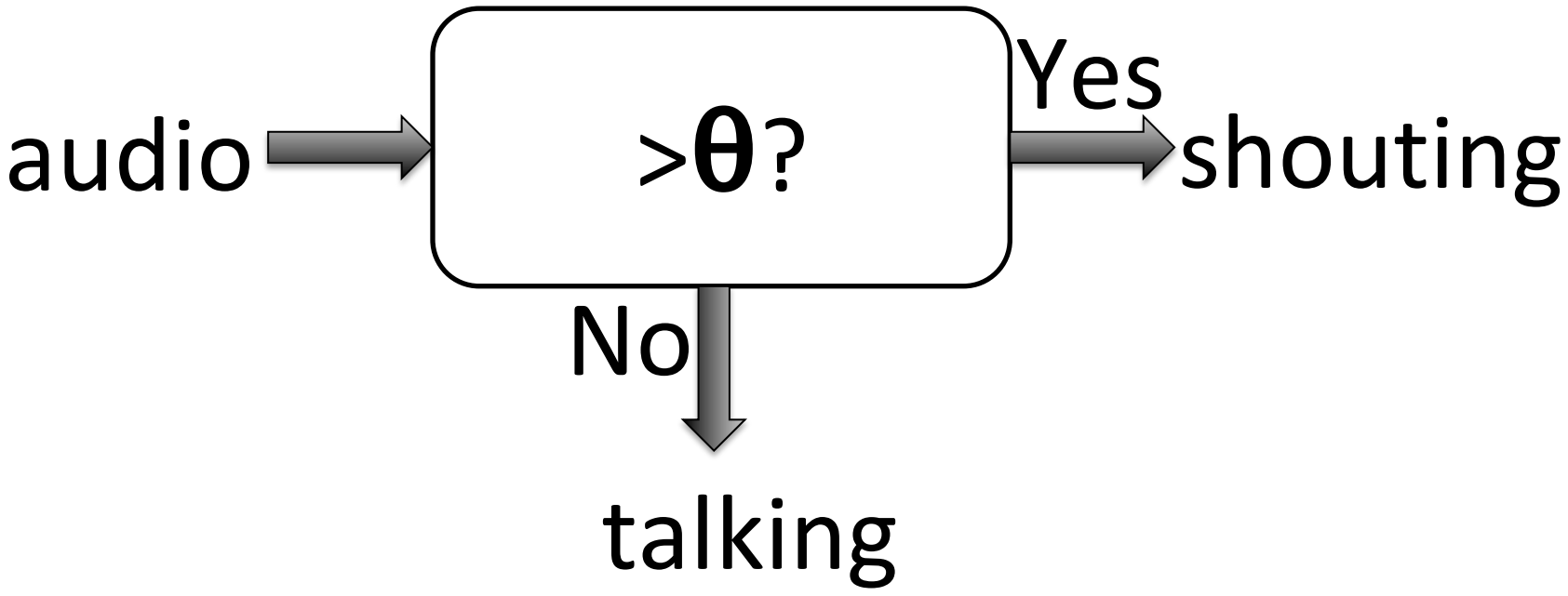
- Detect semantics behind the data
  - Is this document a research article or a love letter?
  - What object is there in the image?
  - Whose voice is that?
  - What activity are people doing in the video?

Generic Analysis Model



You have to tell the processor how to interpret the data!

Example: thresholds



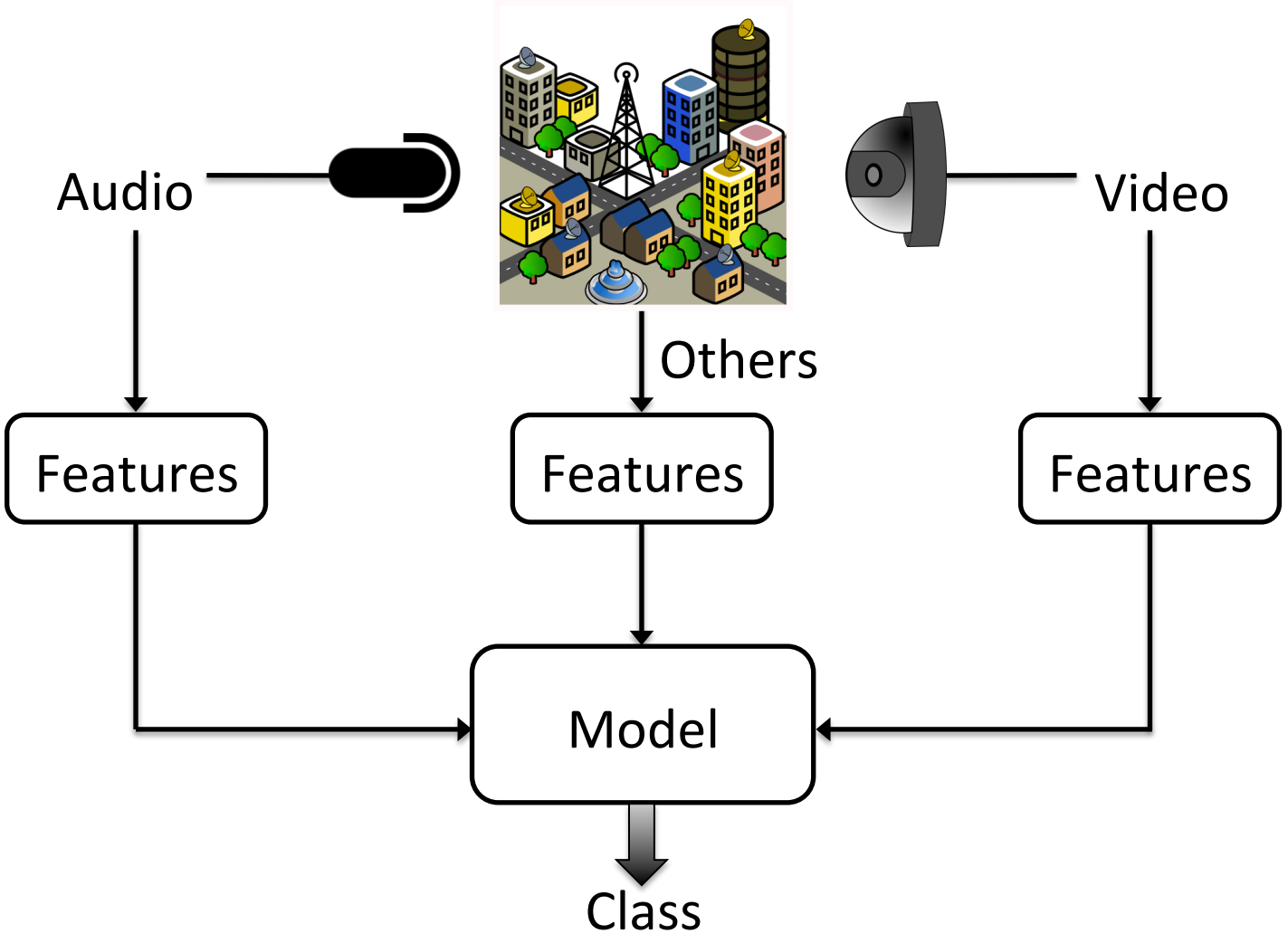
How to determine the threshold?

Main Problems

1. The number of samples can be very large!
2. The number of thresholds (features) can be very large!
3. Visualization may be difficult!

Machine learning tells the processor how to “interpret” new data by “learning” from the past experience!

Machine Learning for Multimedia



A Basic Learner

- Store samples
- Compare input with the stored samples
- Apply label of closest sample

f^a_{11}	f^a_{12}	f^v_{13}	f^v_{14}	f^v_{15}	f^t_{16}	1 (Shouting)
f^a_{21}	f^a_{22}	f^v_{23}	f^v_{24}	f^v_{25}	f^t_{26}	0 (Talking)
f^a_{31}	f^a_{32}	f^v_{33}	f^v_{34}	f^v_{35}	f^t_{36}	1 Shouting
…	…	…	…	…	…	…

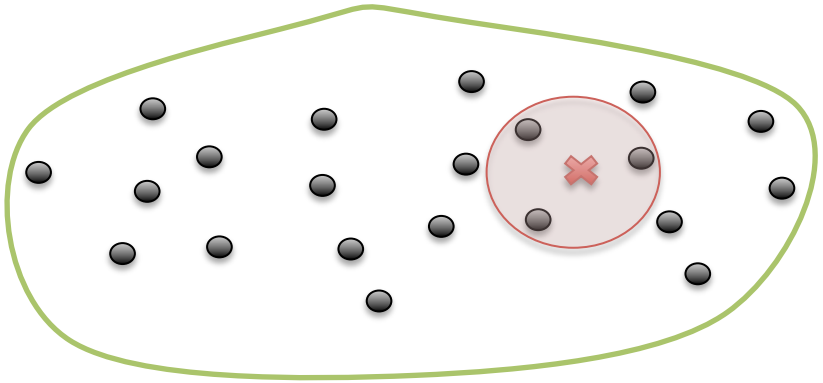
A Basic Learner

- Store samples
- Compare input with the stored samples
- Apply label of closest sample

Problems: Noise.
Relying on single sample is not robust!

K-Nearest Neighbor (KNN)

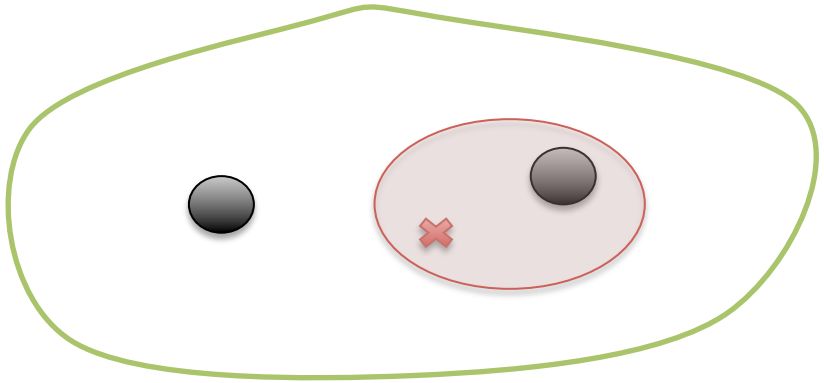
- Find K-Nearest samples!
- Choose label that occurs most!



Problems
1. High memory requirements!
2. Inefficient!

Parameterized Technique

- One representative sample for each class
  - similar to k-Means
- Memory = number of classes
- Efficient



Machine Learning Steps

1. Collect data Samples
2. Label the Samples
3. Choose a *Model/Function/Procedure*
4. Determine Parameters
5. Evaluate

Machine Learning Steps

1. Collect data Samples
2. Label the Samples
3. Choose a *Model/Function/Procedure*
4. Determine Parameters
5. Evaluate

Data collection

- All sensor values, media, and textual data
- Calculate features
- At least 50 samples, the more the better!

f^a_{11}	f^a_{12}	f^v_{13}	f^v_{14}	f^v_{15}	f^t_{16}
f^a_{21}	f^a_{22}	f^v_{23}	f^v_{24}	f^v_{25}	f^t_{26}
f^a_{31}	f^a_{32}	f^v_{33}	f^v_{34}	f^v_{35}	f^t_{36}
…	…	…	…	…	…

What features to choose?

Video/Image Features

SIFT	Feature point based
SURF	Feature point based
HOG	Gradient based
Histogram	Color or gray scale
Saliency	Edge, contrast, and intensity
Edge density	Edge based
Image Moments	Weighted average
Motion vectors	Object motion or camera motion

Audio Features

MLCC	Cepstral coefficient
FFT	Frequency components
Filter bank	Frequency components
Pitch	Dominant Frequency
Energy	Loudness
ZCR	Amplitude Based

Textual Features

Bag of words	Word frequency
Happy words	Emotions
Sad words	Emotions
Metadata	Description
Motivational words	Emotions
ASR	Speech recognition

Feature Granularity

- A patch, image, or set of images?
- Audio window size?
- How frequently?

Feature Synchronization

- Continuous media
- Multiple sources, multiple sampling rates
- What constitutes a sample?
- Finest granularity of coarsest granularity?

Machine Learning Steps

1. Collect data Samples
2. Label the Samples
3. Choose a *Model/Function/Procedure*
4. Determine Parameters
5. Evaluate

How to label the data?

- Manual
- User study
- Crowd sourcing

f^a_{11}	f^a_{12}	f^v_{13}	f^v_{14}	f^v_{15}	f^t_{16}	1 (Shouting)
f^a_{21}	f^a_{22}	f^v_{23}	f^v_{24}	f^v_{25}	f^t_{26}	0 (Talking)
f^a_{31}	f^a_{32}	f^v_{33}	f^v_{34}	f^v_{35}	f^t_{36}	1 Shouting
…	…	…	…	…	…	…

Machine Learning Steps

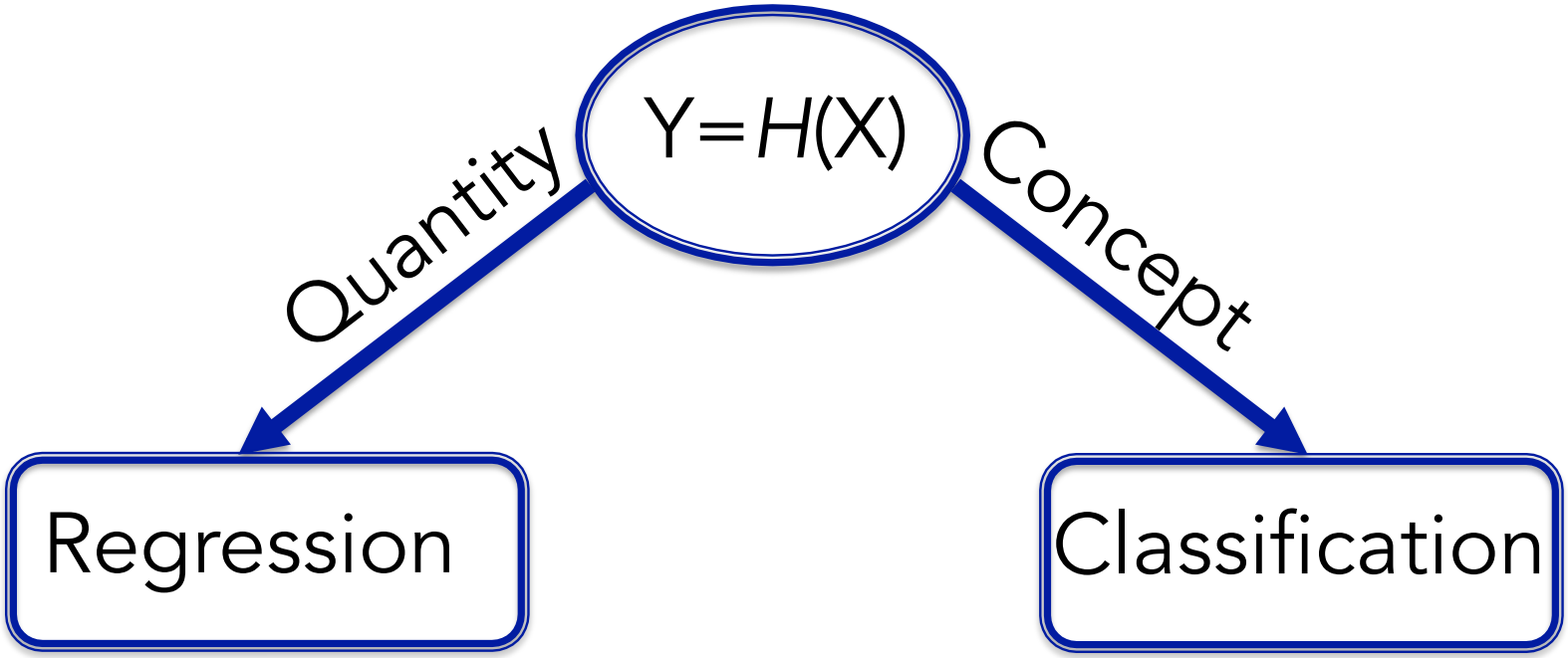
1. Collect data Samples
2. Label the Samples
3. Choose a *Model/Function/Procedure*
4. Determine Parameters
5. Evaluate

Model/Function/Procedure

```
if
   energy>threshold
then
   shouting
else
    talking
end
```

Hypothesis!

What is your output?



Ranking
Value
…

Fight detection
Person detection
…

Regression Methods

- Linear Regression
- Non-linear Regression
- Ordinal Regression
- E.g.

$Y = c+\omega_1 x_1+\omega_2 x_2+\ldots$

Goal: Determine weights to minimize the error!

Classification Methods can be divided based on many criteria!

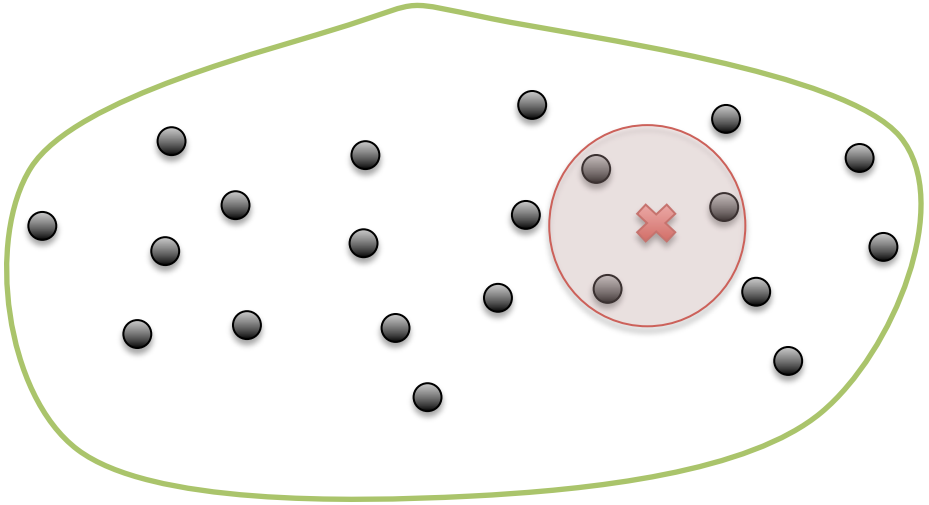
- Parametric/Non-parametric
- Learning approach (incremental/batch)

Non-parametric Methods

- No standard distribution
  - No parameters
- Learner data itself
- Example: kNN

Non-parametric Methods

k-NN
label = majority voting

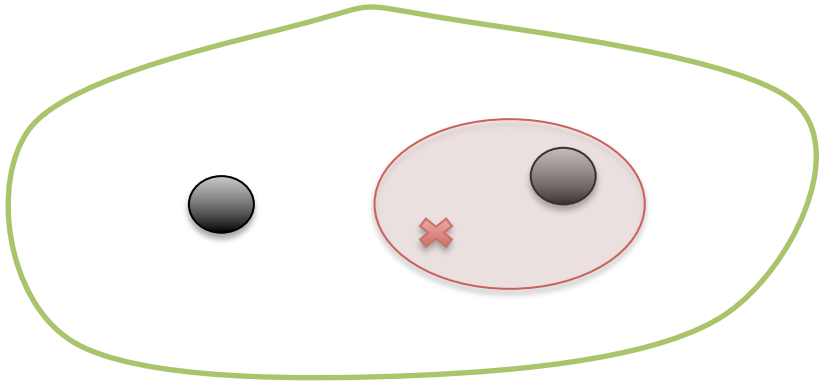


Parametric Methods

- Finite set of parameters
- Do not grow with data samples
- Learning is time consuming, but classification is fast

Parameterized Technique

- One representative sample for each class
  - similar to k-Means
- Memory = number of classes
- Efficient



Parameters

- Probabilities
- Standard distribution parameters
- Weights
- Discriminant coefficients

Will there be food today?

Gender (G)	Supervisor (S)	Food (F)	Count	Probability
Male	Present	Food	10	0.1
		No food	20	0.20
	Absent	Food	3	0.03
		No food	15	0.15
Female	Present	Food	25	0.25
		No food	5	0.05
	Absent	Food	6	0.06
		No food	16	0.16

P(Food)= 0.1 + 0.03 + 0.25 + 0.06 = 0.44

Will there be food today?

Gender (G)	Supervisor (S)	Food (F)	Count	Probability
Male	Present	Food	10	0.1
		No food	20	0.20
	Absent	Food	3	0.03
		No food	15	0.15
Female	Present	Food	25	0.25
		No food	5	0.05
	Absent	Food	6	0.06
		No food	16	0.16

P(Food/Female)= (0.25 + 0.06)/(0.25+0.05+0.06+0.16) = 0.59

Will there be food today?

Gender (G)	Supervisor (S)	Food (F)	Count	Probability
Male	Present	Food	10	0.1
		No food	20	0.20
	Absent	Food	3	0.03
		No food	15	0.15
Female	Present	Food	25	0.25
		No food	5	0.05
	Absent	Food	6	0.06
		No food	16	0.16

P(Food/Female, Present)= 0.25/(0.25+0.05) = 0.83

Summary

With joint probability distribution, we can answer any query about the variables!

Unknown (Food) – latent variable
Evidences or observed variables – Gender, Professor

Probabilistic Classifier

$$L = \underset{k}{\operatorname{argmax}}\ P(C_k/F_i)$$

$$F_i = \{f_1, f_2, f_3\}_i$$

Too many parameters!

$$2^n$$

Learner

x1	x2	x3	P(Y0)	P(Y1)
0	0	0		
0	0	1		
0	1	0		
0	1	1		
1	0	0		
1	0	1		
1	1	0		
1	1	1		

Bayes' Rule

$$P(C/F) = \frac{P(F/C)P(C)}{P(F)}$$

- Parameters needed for P(F) ~ 0
- For P(C) ~ 1
- For $P(F/C)=P(f_1, f_2, f_3/C) \sim 2^n$

Naïve Bayes

Assume x1, x2, and x3 to be conditionally independent!

$$P(f_1, f_2, f_3/C) = P(f_1/C)\, P(f_2/C)\, P(f_3/C)$$

$$P(F/C) \sim 2^n => 2n$$

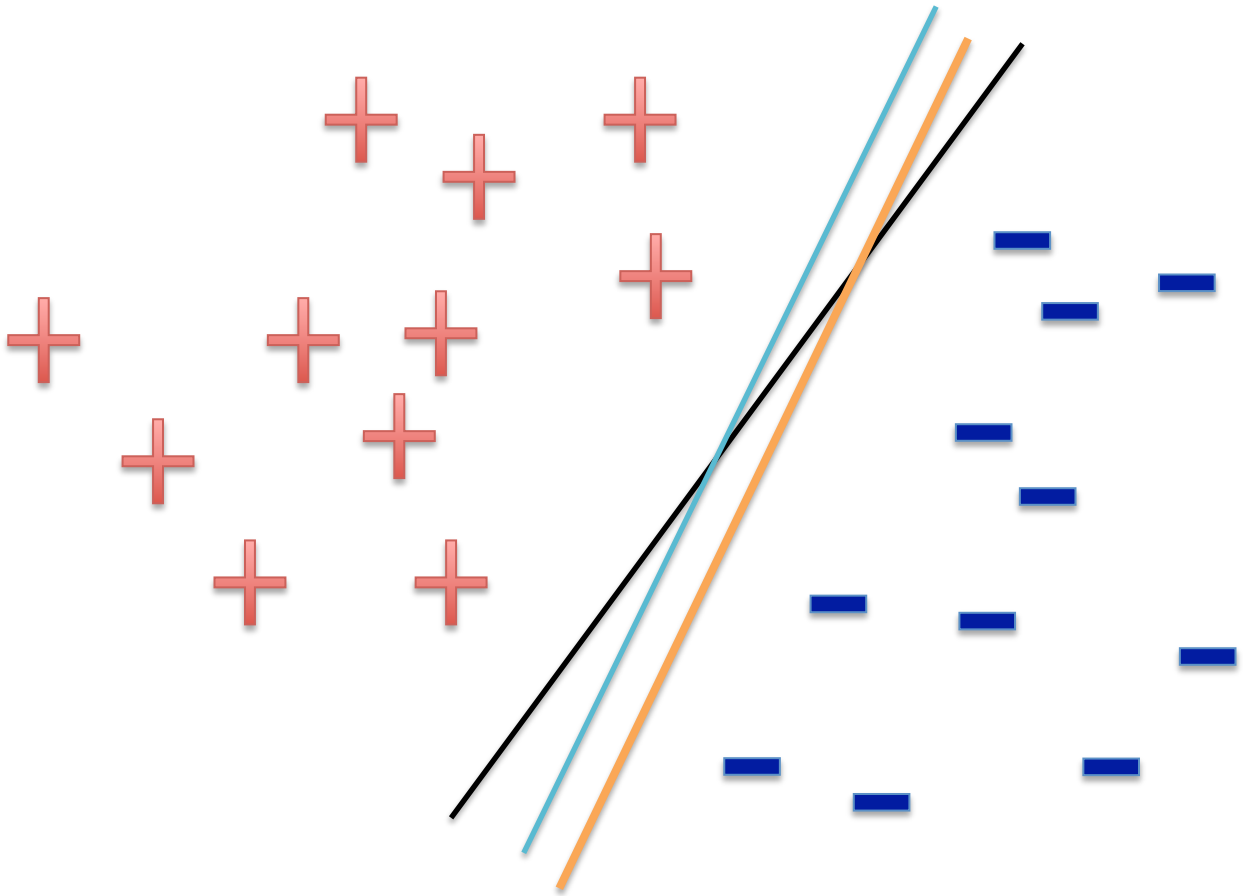
$$L = \underset{k}{\operatorname{argmax}}\ P(C_k/F_i)$$

Non-probabilistic Classifiers

- For large number of features
- Large number of samples
- Example: SVM

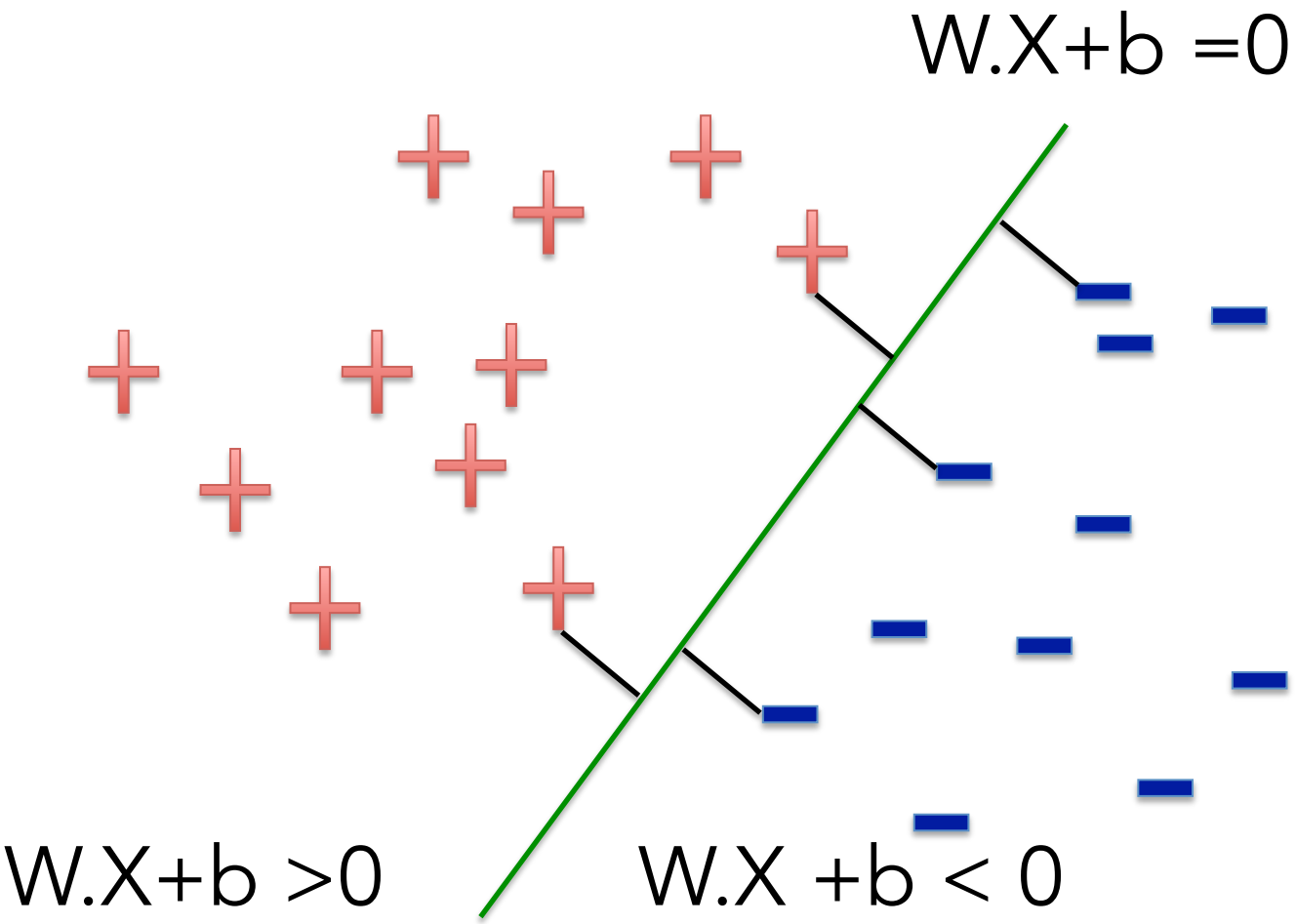
Linear Classifier

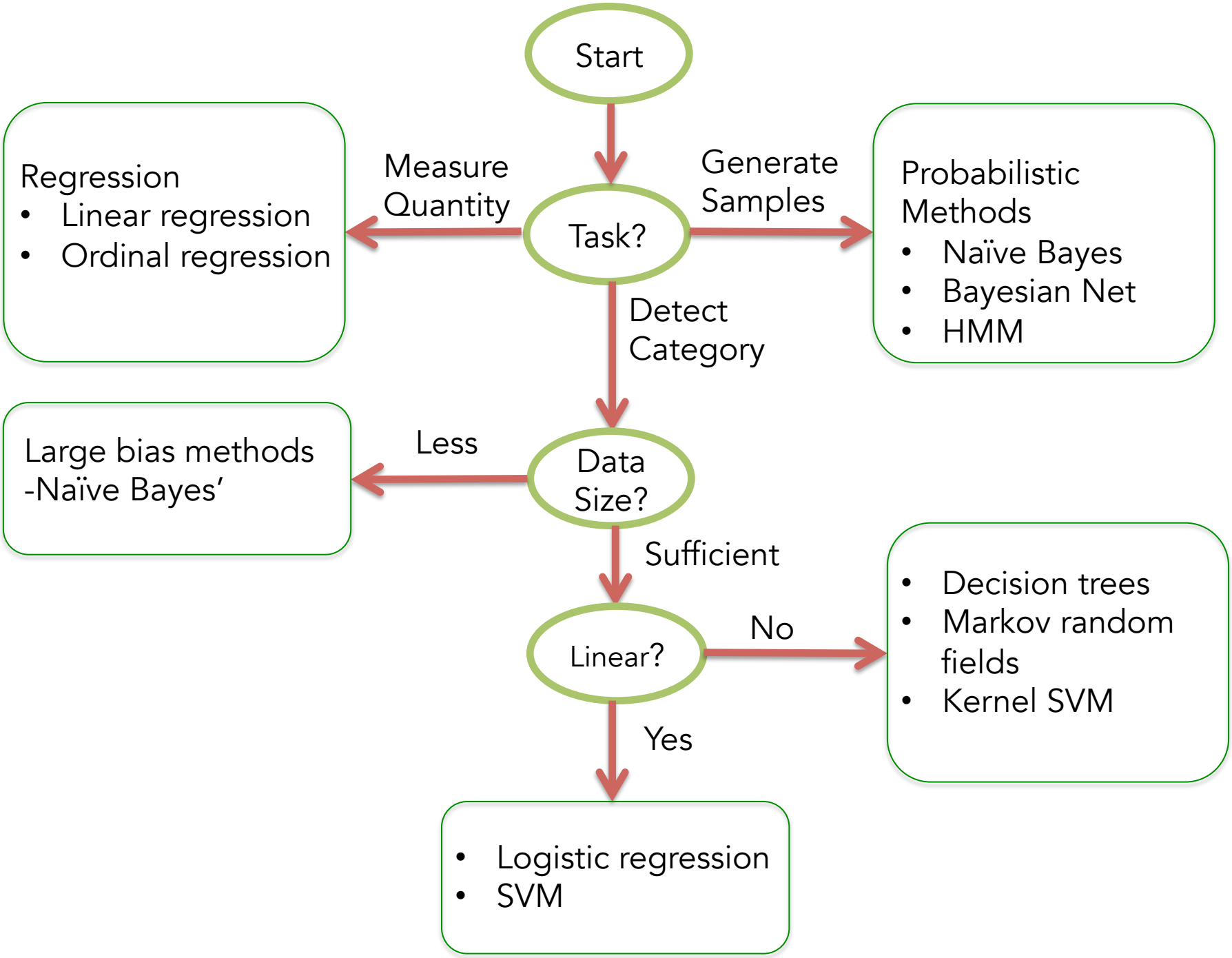
$W.X+b = 0$



Support Vector Machine

Support Vector Machines




Start
Task?
Measure Quantity
Regression
• Linear regression
• Ordinal regression
Generate Samples
Probabilistic Methods
• Naïve Bayes
• Bayesian Net
• HMM
Detect Category
Data Size?
Less
Large bias methods
-Naïve Bayes'
Sufficient
Linear?
No
• Decision trees
• Markov random fields
• Kernel SVM
Yes
• Logistic regression
• SVM

Speed and Accuracy

- Speed
  - Fast: SVM, Naïve Bayes, Logistic Regression, ANN, Parametric, Linear
  - Slow: KNN, Unsupervised, Non-parametric
- Accuracy
  - Try different methods

Ease of Making Online

- Probabilistic methods are generally easy to train online, e.g., Bayesian methods and Logistic Regression
- Discriminant based methods are generally not suitable for online learning, e.g., SVM and decision tree

Temporal Pattern/Class

- HMM
- DNB
- And a billion variants of HMM

Number of features

- For large number of features, go for linear classifiers
  - SVM outperforms both Naïve Bayes and Logistic Regression
- Small number of features (but large number of samples)
  - Decision trees, Kernel SVM

Machine Learning Steps

1. Collect data Samples
2. Label the Samples
3. Choose a *Model/Function/Procedure*
4. Determine Parameters
5. Evaluate

Different Ways to Learn Parameters

- Batch learning
- Online learning
- Reinforcement learning

Machine Learning Steps

1. Collect data Samples
2. Label the Samples
3. Choose a *Model/Function/Procedure*
4. Determine Parameters
5. Evaluate

Conventional Validation

- Train using ~70 % of the data
- Test using ~30 % of the data

What if you do not have too many samples?

Cross Validation

- Leave-p-Out
- K-Fold validation
- Final parameters
  - Lowest error
  - or whole dataset

Performance Measure

- Confusion matrix
- Classification accuracy
- Weighted Accuracy
- Precision/Recall

Confusion Matrix

	Predicted 0	Predicted 1
True 0	44	6
True 1	7	43

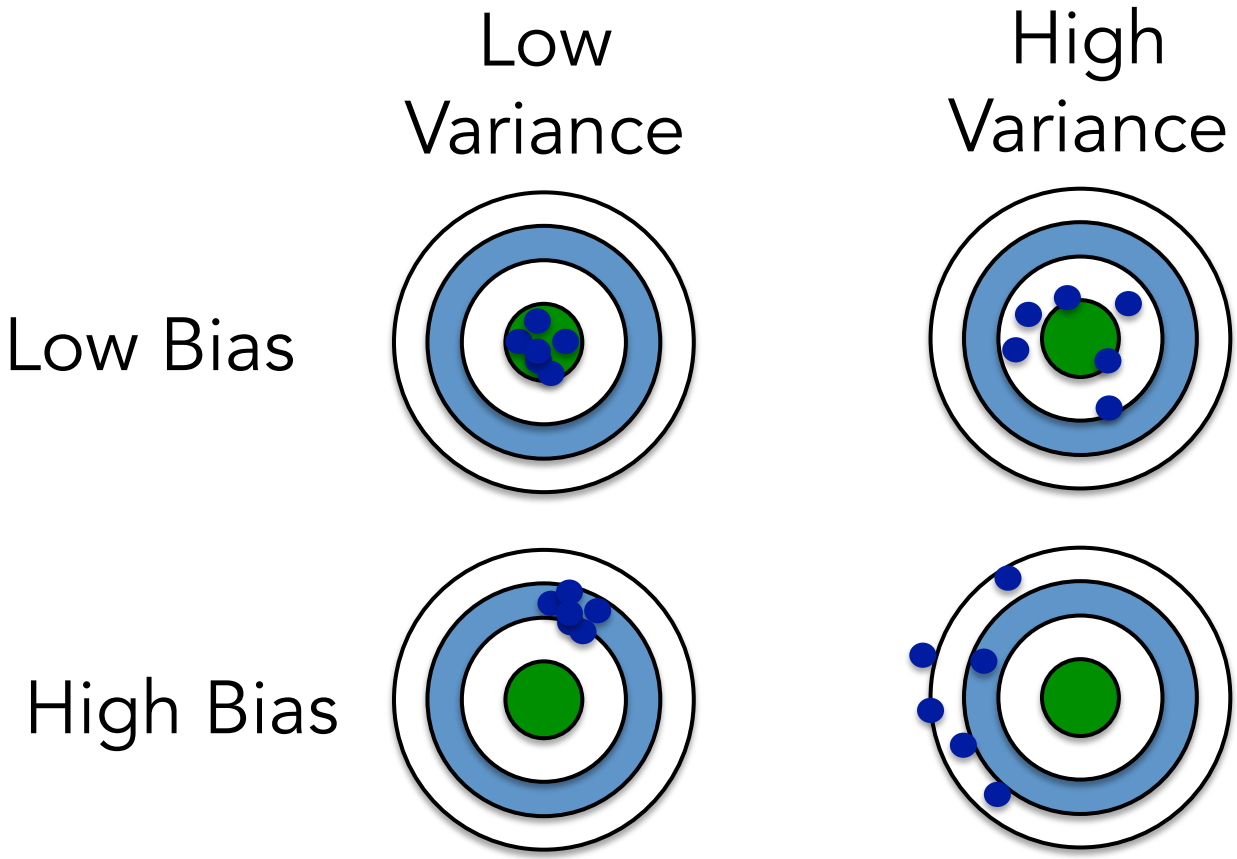
What to try next?

- Get more training data
- Try a smaller set a features
- Try getting additional features
- Adding polynomial features
- Building your own, new, better features
- Bias variance tradeoff

Model complexity Vs Bias/Variance

- Complex models => Low Bias, high variance, good for large size data
- Simple models => High Bias, low variance, good for small size data
- High Bias means underfit while High variance means overfit!
- It is better to underfit when the data size is small.

Bias-Variance Graphical View



Thank you

Essentially, all models are wrong, but some are useful!

-George E. P.