

Discriminative Learning, Validation, & Bayesian Thinking

Alexander G. Ororbia II

Introduction to Machine Learning

CSCI-335

11/4/2026

Special Thanks: Sargur N. Srihari, Andrew Ng, Carlos Guestrin

Multinoulli Regressor: Architecture

We will learn parameters by minimizing the negative log likelihood of our model's predictive posterior. If we say that $\mathbf{p}$ is our model's vector of normalized output probabilities, we define the negative log loss (cost) as follows:

$$\mathbf{p}_k = \frac{e^{\mathbf{f}_k}}{\sum_j e^{\mathbf{f}_j}}, \qquad \mathcal{J} = -\frac{1}{N}\sum_i \Big(\log(\mathbf{p}_{y_i})\Big) + \frac{\lambda}{2}\sum_d \sum_k (W_{d,k})^2$$

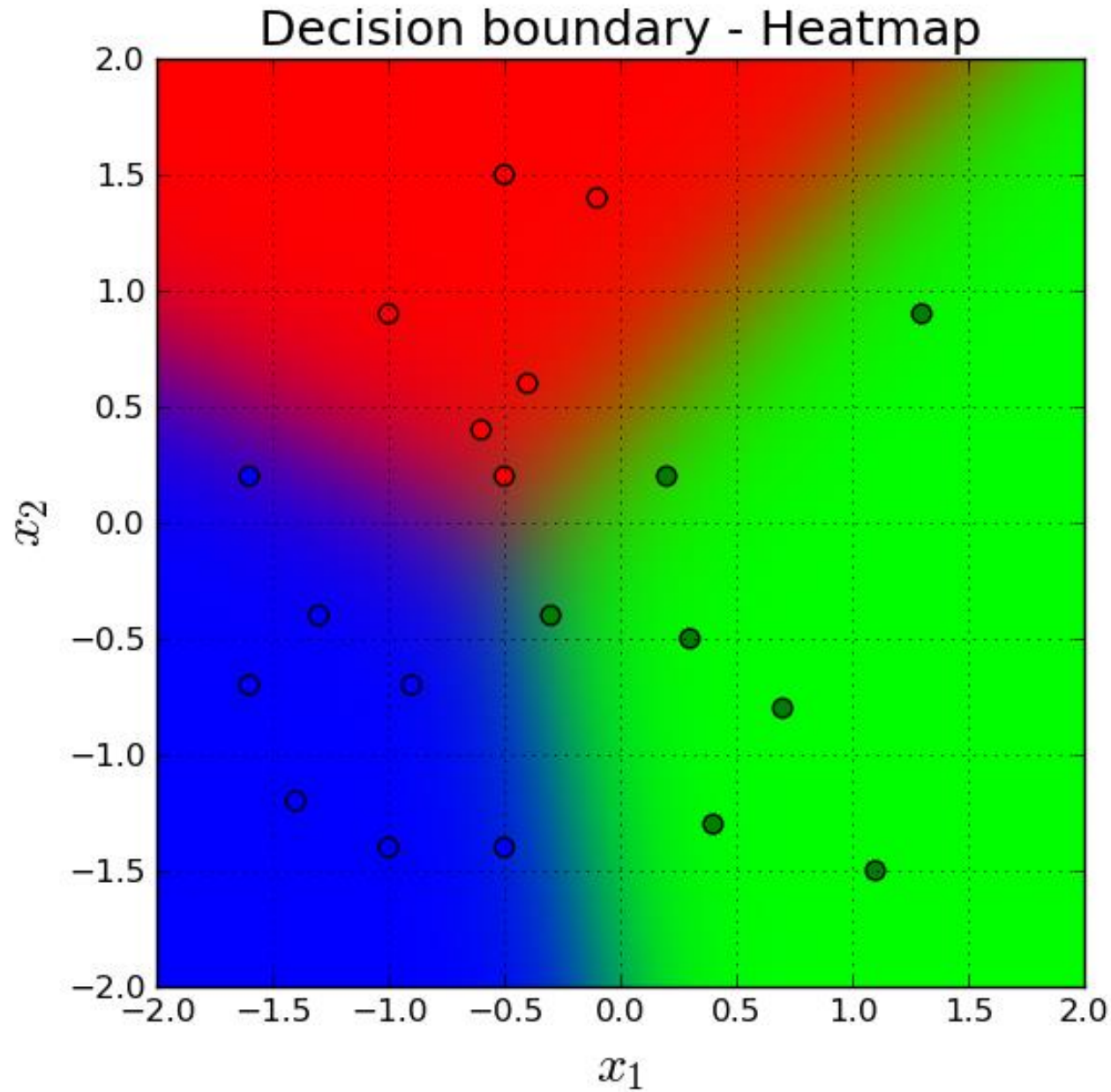
```
Inputs X:
 [[ 0.1  0.5]
 [ 1.1  2.3]
 [-1.1 -2.3]
 [-1.5 -2.5]]

[[ 1.  0.  0.]
 [ 0.  1.  0.]
 [ 0.  0.  1.]
 [ 0.  0.  1.]]
```

$$\frac{\partial \mathcal{J}_i}{\partial \mathbf{f}_k} = \mathbf{p}_k - \mathbf{1}_{y_i=k}$$

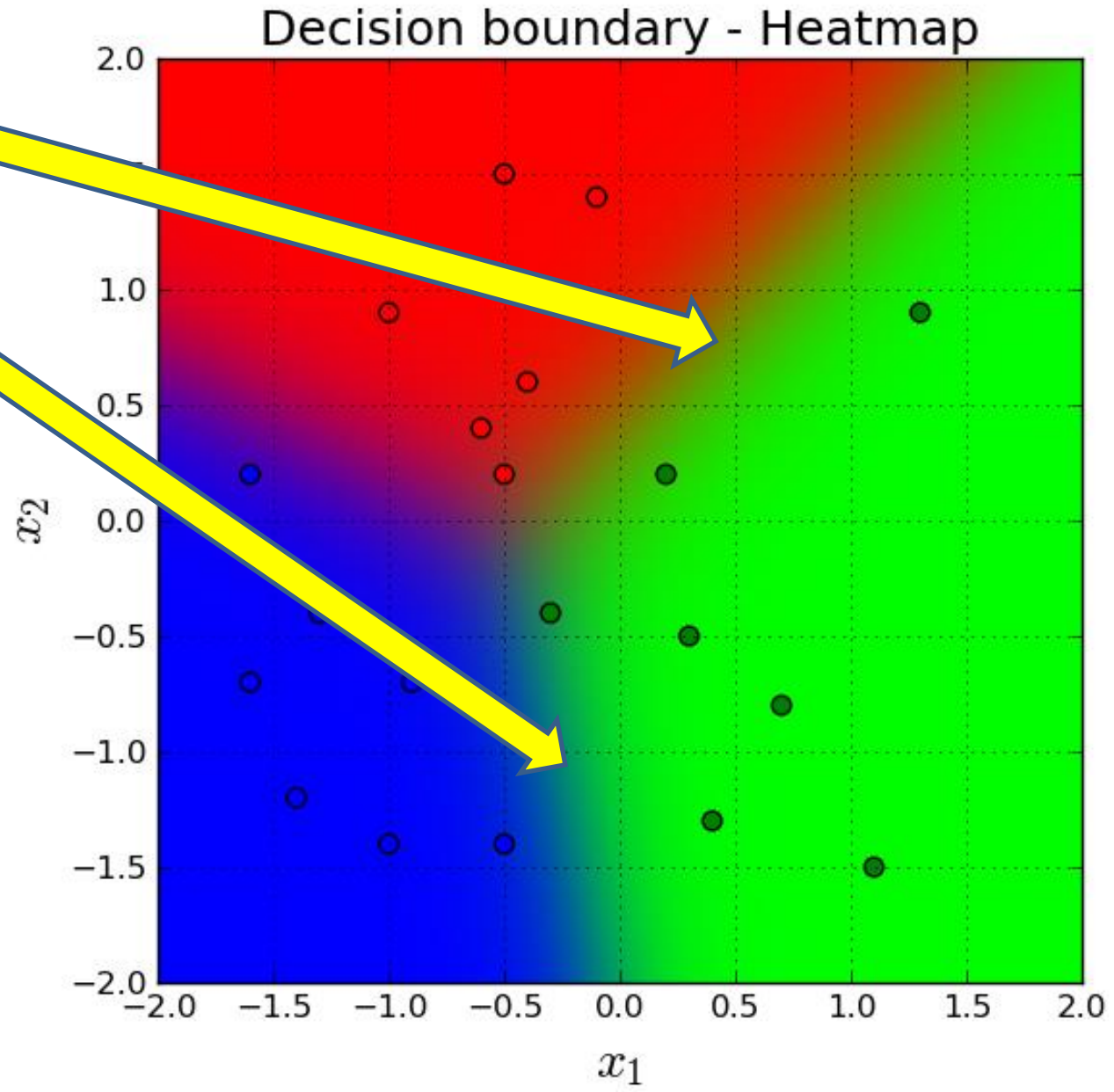
Recall: what is this term known as?

$$\frac{\partial \mathcal{J}}{\partial W} = \mathbf{X}^T * \left(\frac{1}{N}\frac{\partial \mathcal{J}}{\partial \mathbf{f}_k}\right) + \lambda(W) = \mathbf{X}^T * \left(\frac{1}{N}(\mathbf{p}_k - \mathbf{1}_{y_i=k})\right) + \lambda(W)$$

$$\frac{\partial \mathcal{J}}{\partial b} = \sum_{n=0}^{N} \frac{1}{N}\frac{\partial \mathcal{J}}{\partial \mathbf{f}_k} = \sum_{n=0}^{N} \frac{1}{N}(\mathbf{p}_k - \mathbf{1}_{y_i=k})$$


Decision boundary - Heatmap
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
x_2
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
x_1

The boundaries



Notice that the data is linearly separable!

Could We Do Even Better?

- Yes, if we revisit artificial features / basis functions…
  - Generalize these to the notion of infinite-dimensional feature spaces (leading to the “kernel’ trick)
  - Incorporate the concept of “support vectors” and max-margin optimization
  - We get ***support vector machines (SVMs; not treated in this course)***
- Probabilistic graphical models (PGMs)
  - We introduce domain knowledge to craft graphs/relationships of variables with probability arcs/tables (we will be transitioning to this)
- Deep learning
  - Through artificial neural networks or chains of nonlinear processing
  - We will get to this later

Thinking about training and validation…

A Cautionary Example



Image classification of tanks. Autofire when an enemy tank is spotted.
Input data: Photos of own and enemy tanks.
Worked really good with the training set used.
In reality, it failed completely. **Why?**

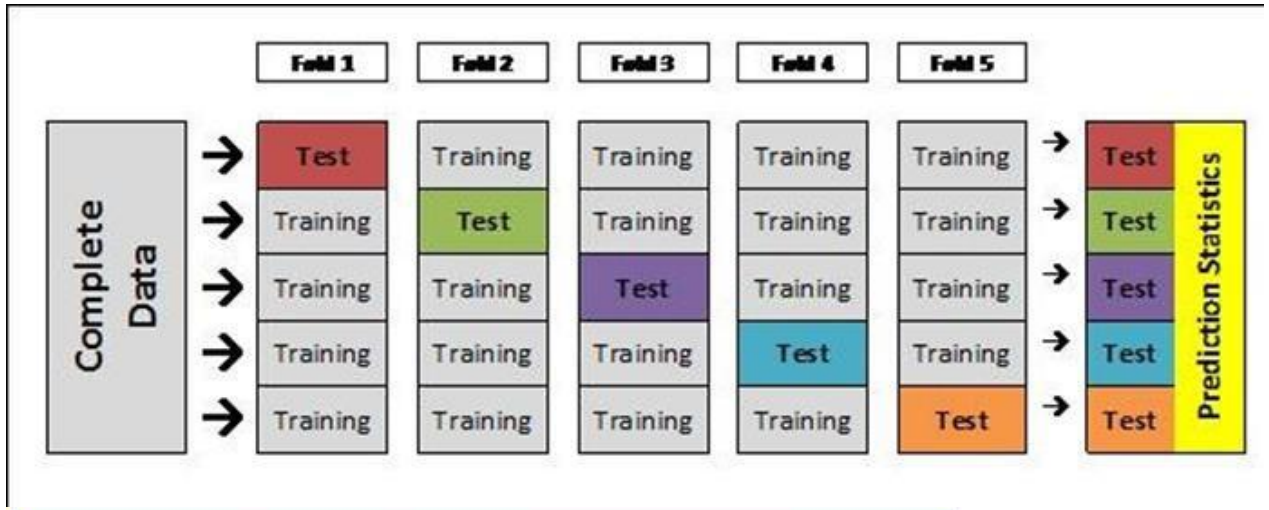
A Cautionary Example

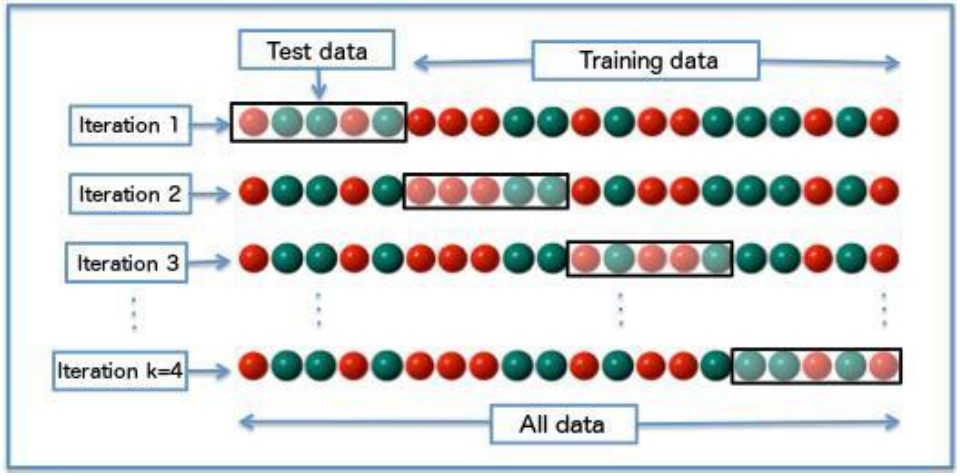


Image classification of tanks. Autofire when an enemy tank is spotted.
Input data: Photos of own and enemy tanks.
Worked really good with the training set used.
In reality, it failed completely.

Reason: All enemy tank photos taken in the morning. All tanks in data shown at dawn
The classifier could NOT recognize dusk from dawn!!!!

Cross-Validation: Measuring Generalization





Good for when dataset is small! Average over trials/folds to get a biased estimator of generalization error.

Generation vs. Discrimination

Generative Models

- Represent both the data and the labels
- Often, makes use of conditional independence and priors
- Examples
  - Naïve Bayes classifier
  - Bayesian network
  - Single/Dual-wing harmonium
  - Variational autoencoder

- Models of data may apply to future prediction problems

Discriminative Models

- Learn to directly predict the labels from the data
- Often, assume a simple boundary (e.g., linear, piecewise)
- Examples
  - Logistic/Multinoulli regression
  - SVM, perceptron, discriminants
  - Decision tree / An ensemble
  - MLP

- Often easier to predict a label from the data than to model the data

Generation vs. Discrimination

Generative Models

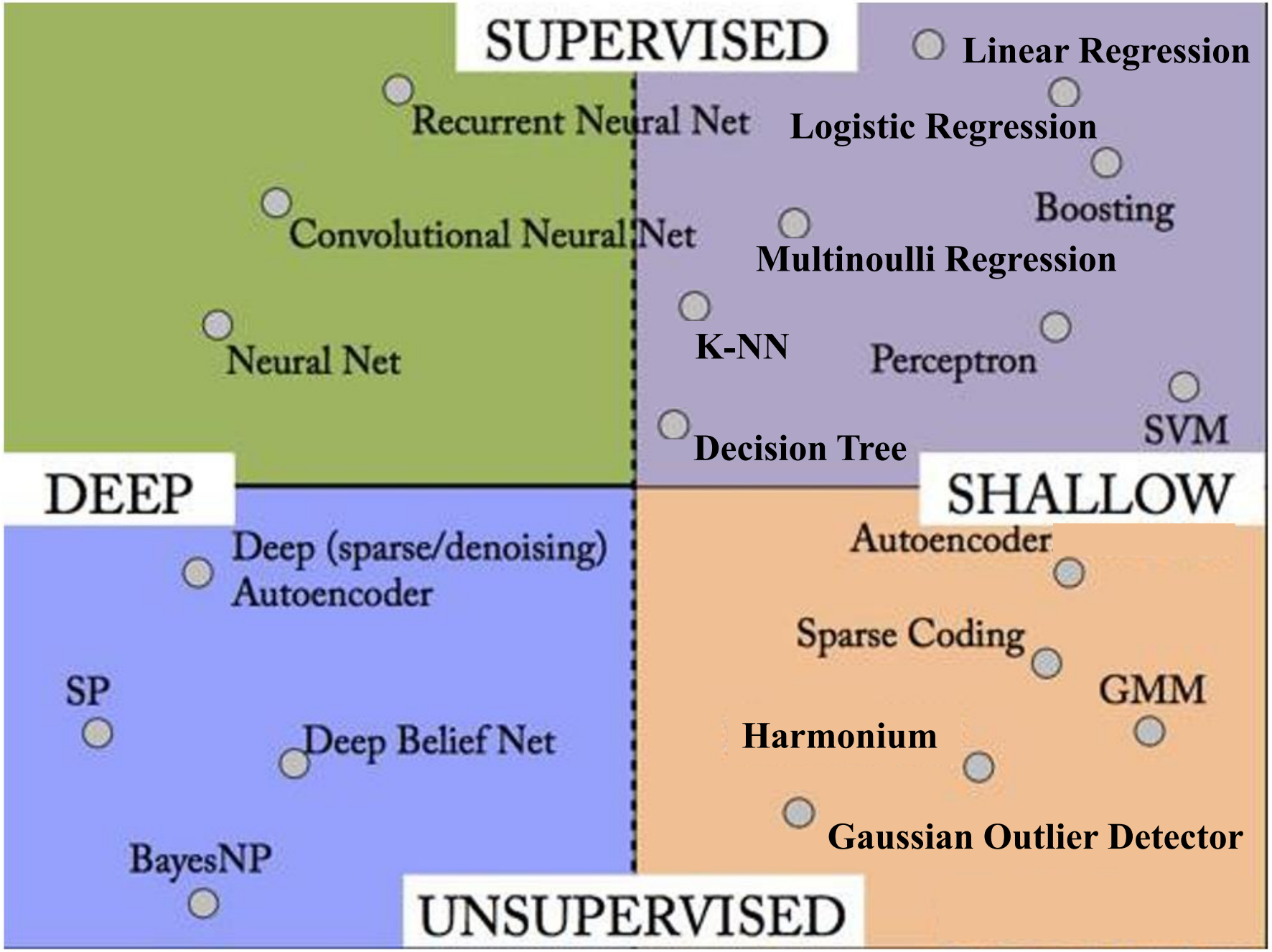
- Represent both the data and the labels
- Often, makes use of conditional independence and priors
- Examples
  - Naïve Bayes classifier
  - Bayesian network
  - Single/Dual-wing harmonium
  - Variational autoencoder

- Models of data may apply to future prediction problems

Discriminative Models

- Learn to directly predict the labels from the data
- Often, assume a simple boundary (e.g., linear, piecewise)
- Examples
  - Logistic/Multinoulli regression
  - SVM, perceptron, discriminants
  - Decision tree / An ensemble
  - MLP

- Often easier to predict a label from the data than to model the data



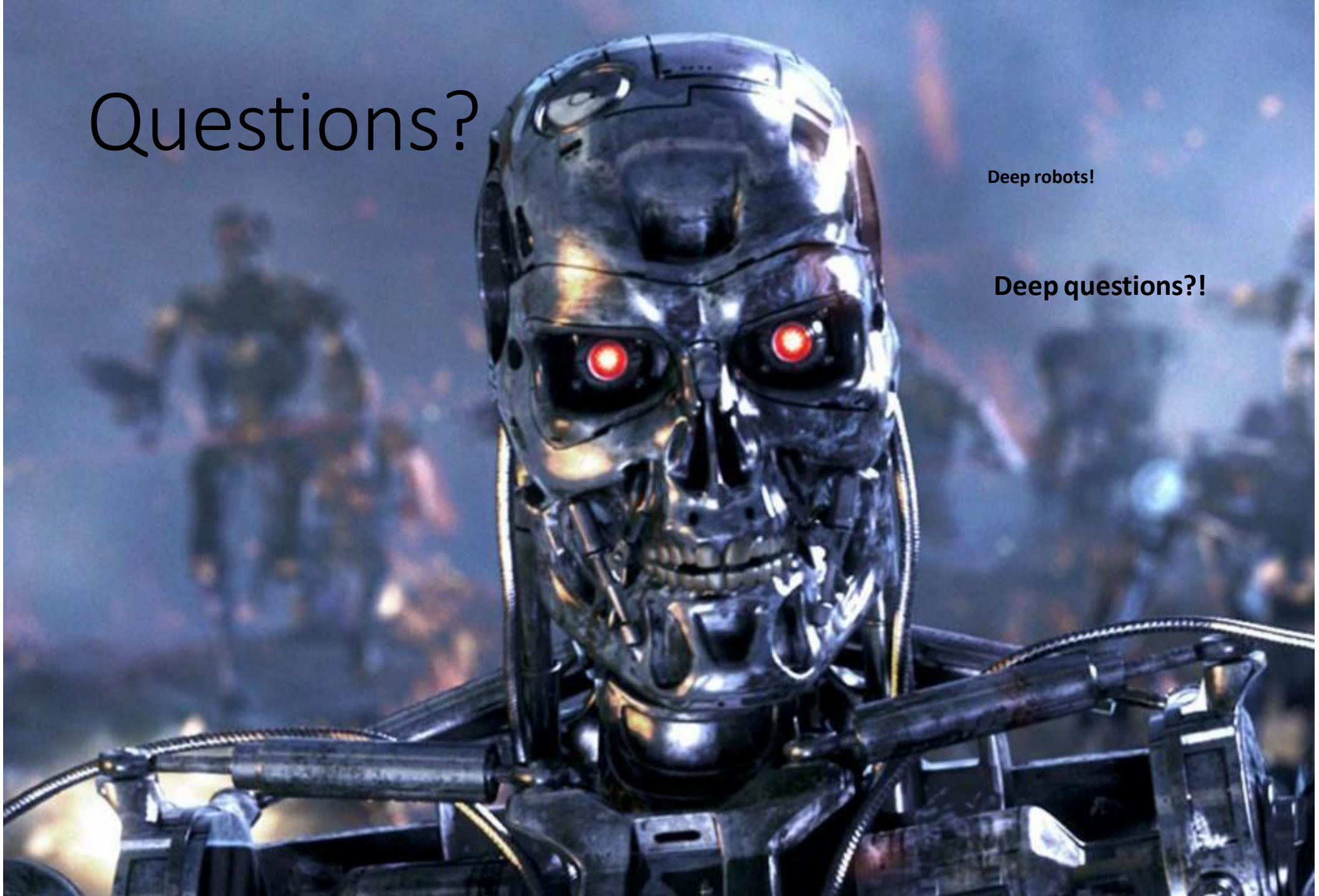
Another machine learning modeling spectrum

Getting Back to Bayes Thinking

Bayes Theorem Revisited!



Questions?



Deep robots!

Deep questions?!