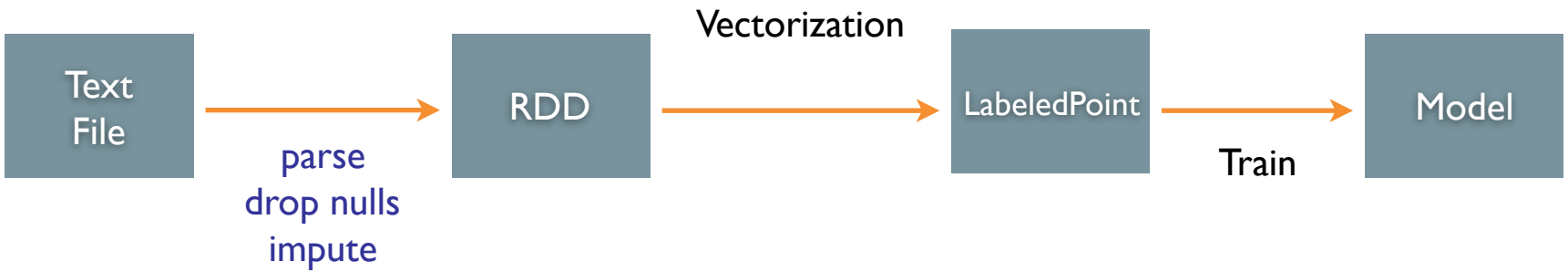
livelessons
video instruction from technology experts

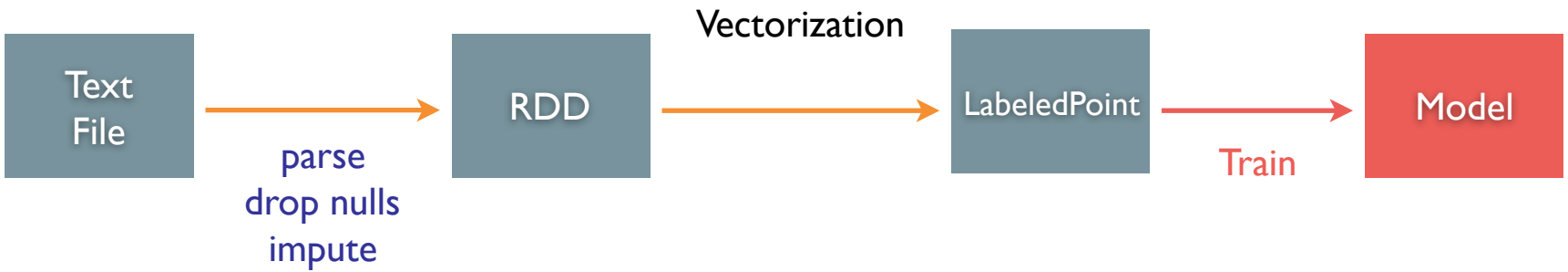
Lesson 5: Advanced Applications

5.3 Introduction to Supervised Learning: Logistic Regression

Machine Learning Pipeline



Machine Learning Pipeline



Supervised Learning

- Training Data includes desired output

Unsupervised Learning

- Training Data does not include desired output

Semi-supervised Learning

- Training Data includes a few desired outputs

Reinforcement Learning

- Rewards from sequence of actions

Classification:

- Spam Filtering and document classification
- Finance: Fraud detection and loan default prediction
- Sentiment Analysis: People like to do this with Tweets
- Customer relationship management: Churn Analysis

Iris Dataset

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	label
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

Features
(feature matrix)

Target

Train

Input: historical labeled data

+

(hypothesis) function with unknown parameter values
<linear, logistic, etc.>

=

Predict

Output: parameter values

Unified Platform

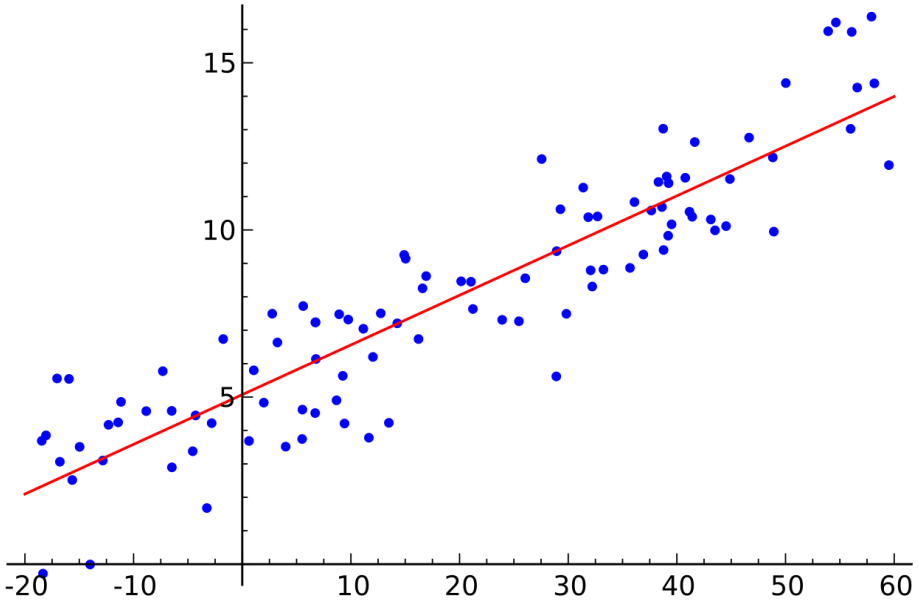
MLlib Supported Methods (1.4.1)

Problem	Method
Binary Classification	*linear SVMs, logistic regression, decision trees, random forests, gradient-boosted trees, naive Bayes*
Multiclass Classification	*logistic regression, decision trees, random forests, naive Bayes*
Regression	*linear least squares, Lasso, ridge regression, decision trees, random forests, gradient-boosted trees, isotonic regression*

MLlib Supported Methods (1.4.1)

Problem	Method
Binary Classification	linear SVMs, logistic regression, decision trees, random forests, gradient-boosted trees, naive Bayes
Multiclass Classification	logistic regression, decision trees, random forests, naive Bayes
Regression	linear least squares, Lasso, ridge regression, decision trees, random forests, gradient-boosted trees, isotonic regression

Linear Regression



Parameters

$$A = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Logistic Regression

Want:

$$0 < P(label \mid X) < 1$$

Have:

$$A = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Logistic Regression

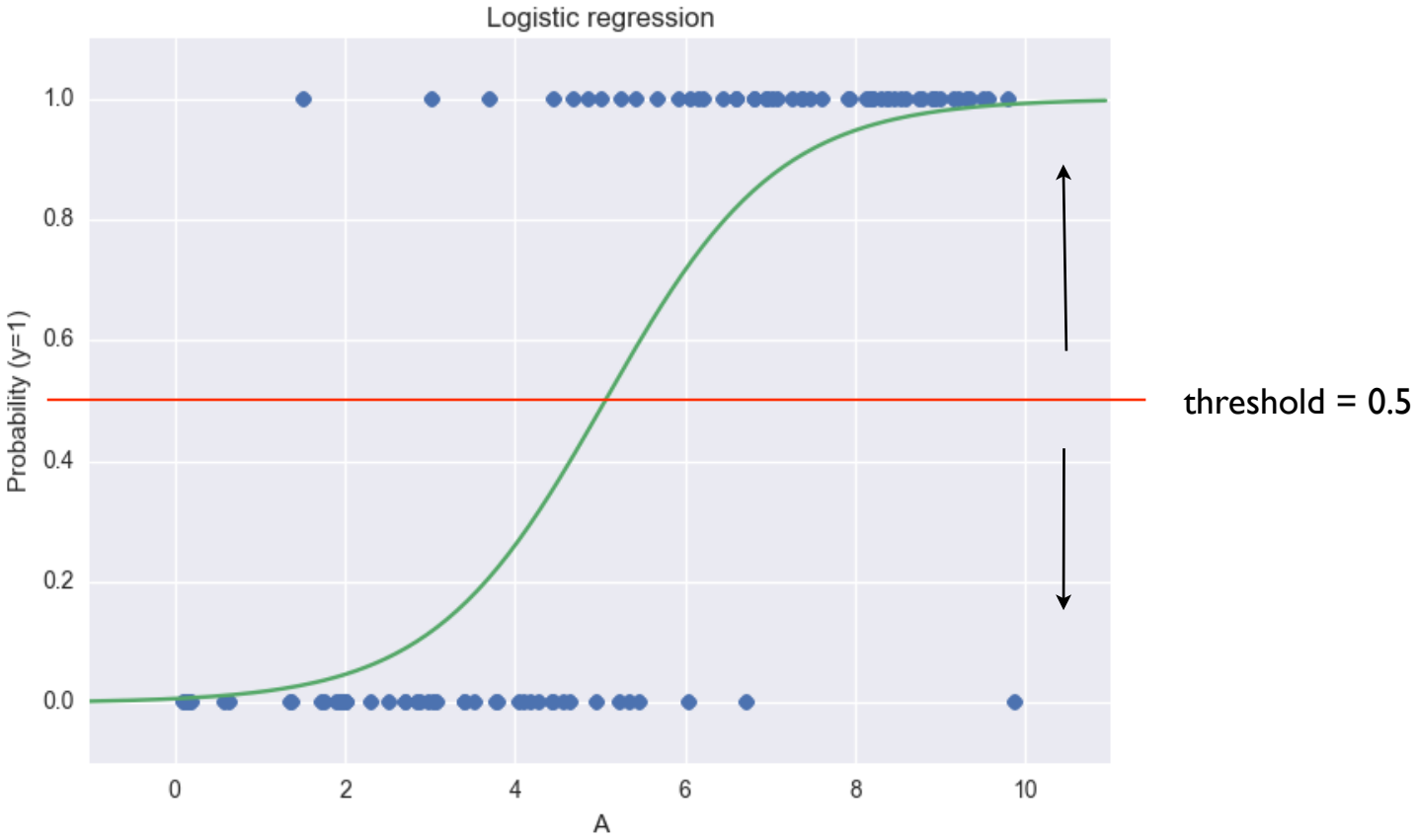
$$A = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

$$P(label \mid X) = \sigma(A)$$

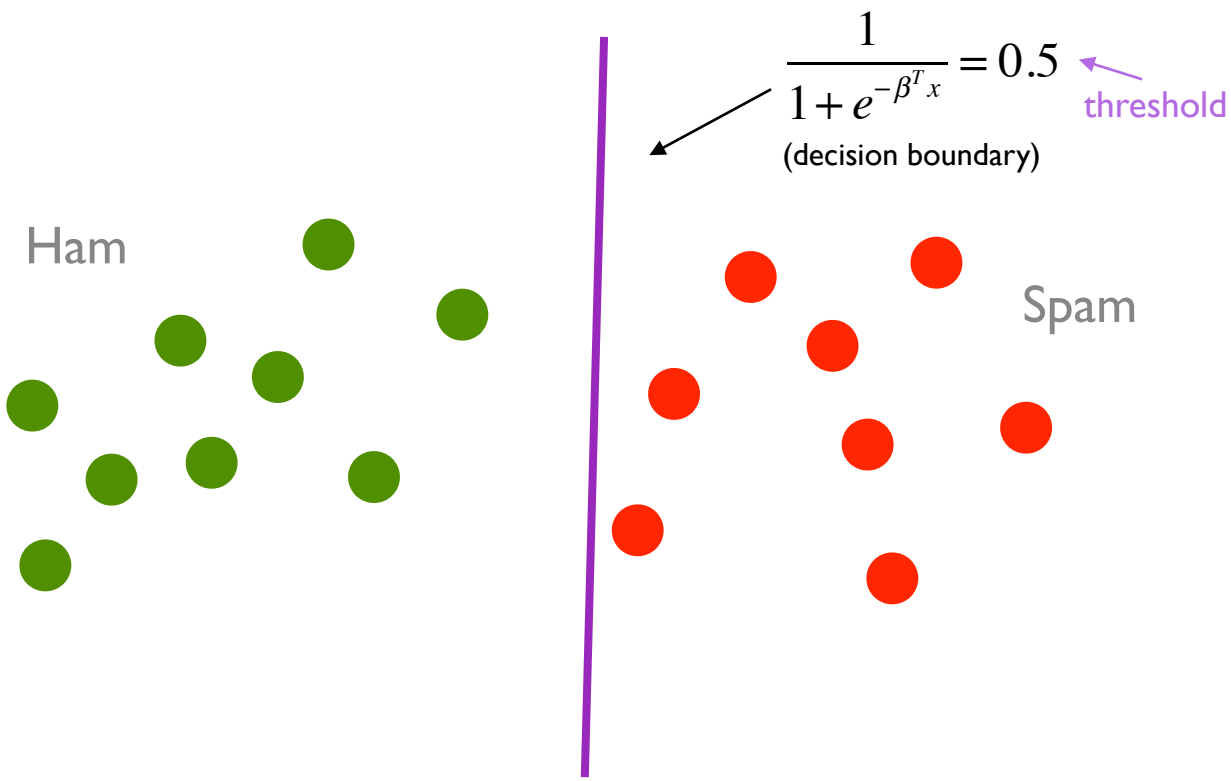
$$\sigma = \frac{1}{1 + e^{-A}}$$ (function bound between 0 and 1)

Logistic Regression

(contrary to its name... actually used to classify)



Linear Separator



Linear Separator