Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 3: Classification overview

Jan 27, 2016

Classification

A mapping *h* from input data x (drawn from instance space $\mathcal{X}$) to a label (or labels) y from some enumerable output space $\mathcal{Y}$

$\mathcal{X}$ = set of all skyscrapers
$\mathcal{Y}$ = {art deco, neo-gothic, modern}

x = the empire state building
y = art deco

Recognizing a Classification Problem

- Can you formulate your question as a *choice* among some universe of possible classes?

- Can you create (or find) labeled data that marks that choice for a bunch of examples? Can *you* make that choice?

- Can you create features that might help in distinguishing those classes?

1. Those that belong to the emperor
2. Embalmed ones
3. Those that are trained
4. Suckling pigs
5. Mermaids (or Sirens)
6. Fabulous ones
7. Stray dogs
8. Those that are included in this classification
9. Those that tremble as if they were mad
10. Innumerable ones
11. Those drawn with a very fine camel hair brush
12. Et cetera
13. Those that have just broken the flower vase
14. Those that, at a distance, resemble flies

The “Celestial Emporium of Benevolent Knowledge” from Borges (1942)

Conceptually, the most interesting aspect of this classification system is that it does not exist. Certain types of categorizations may appear in the imagination of poets, but they are never found in the practical or linguistic classes of organisms or of man-made objects used by any of the cultures of the world.

Eleanor Rosch (1978),
“Principles of Categorization”

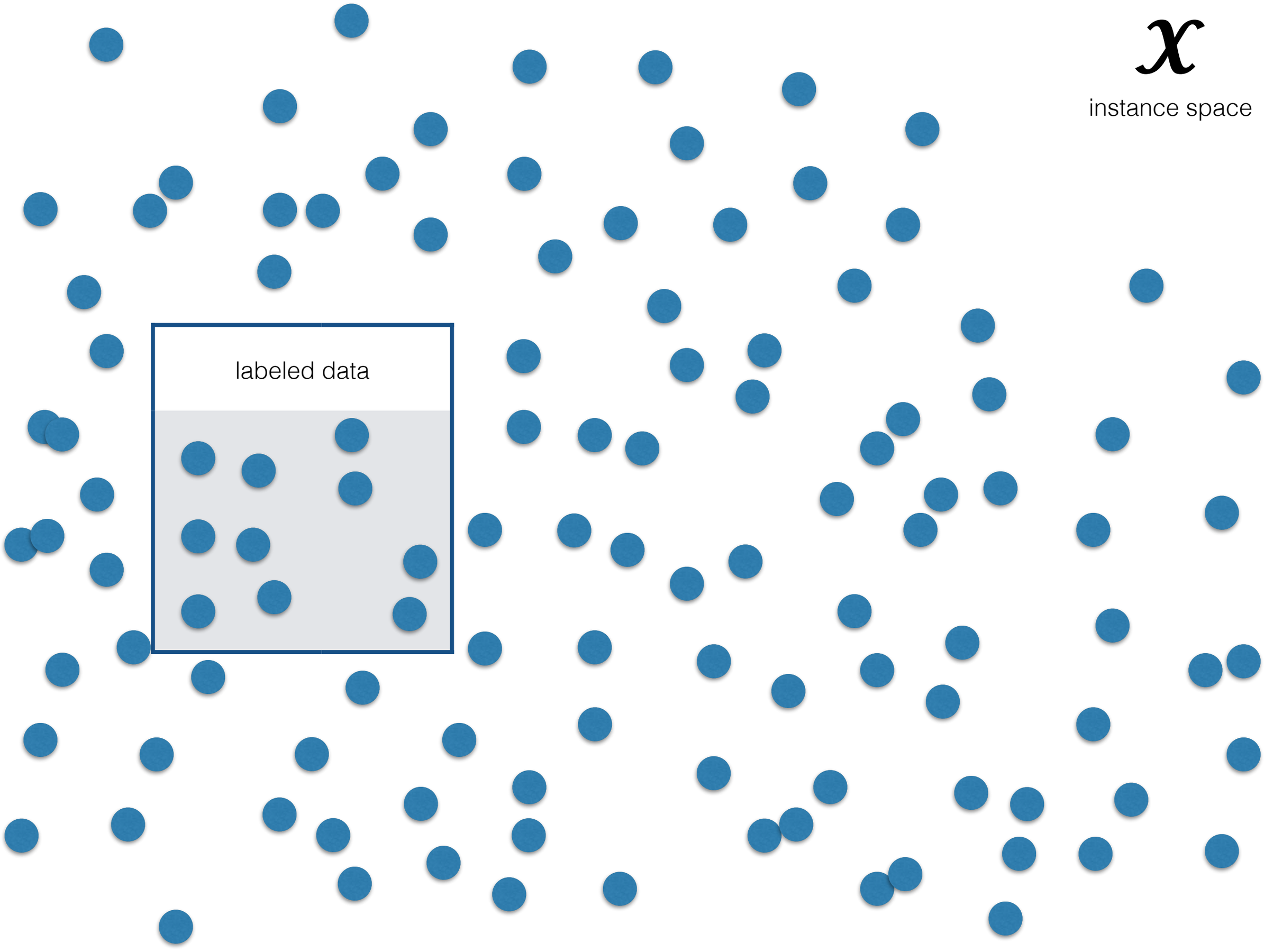
Evaluation

- For all supervised problems, it's important to understand how well your model is performing

- What we try to estimate is how well you will perform in the future, on new data also drawn from $\boldsymbol{x}$

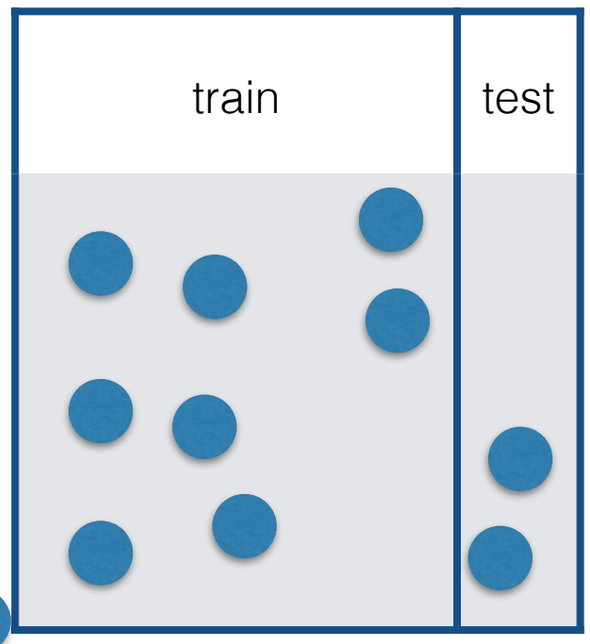
- Trouble arises when the training data <x, y> you have does not characterize the full instance space.

    - n is small
    - sampling bias in the selection of <x, y>
    - x is dependent on time
    - y is dependent on time (concept drift)


$\mathcal{X}$
instance space
labeled data

$\mathcal{X}$

instance space

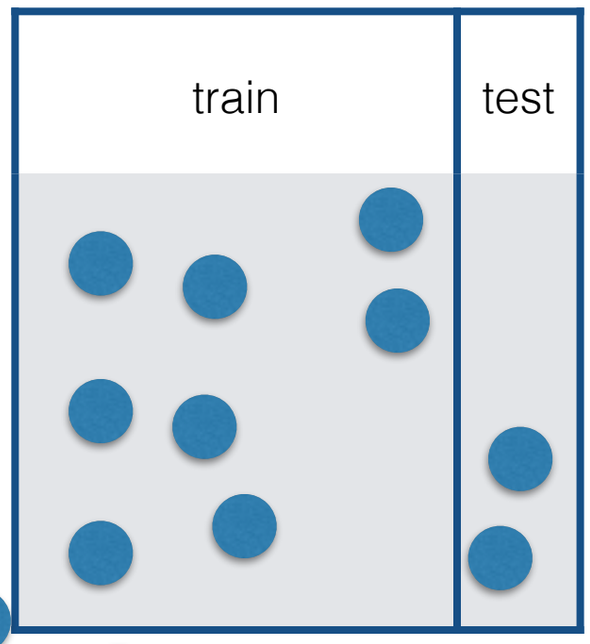


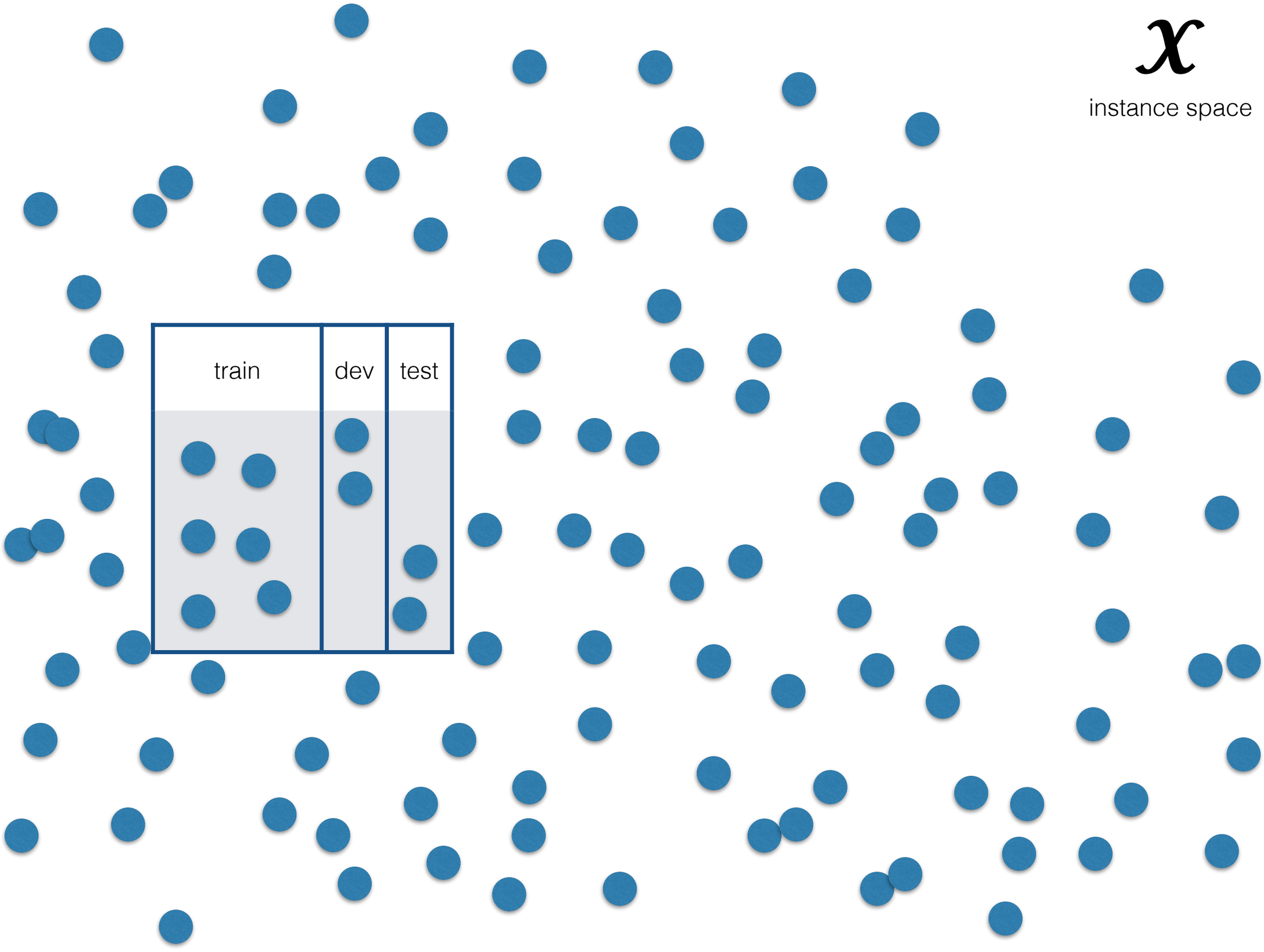
Train/Test split

- To estimate performance on future unseen data, train a model on 80% and test that trained model on the remaining 20%
- What can go wrong here?

$\mathcal{X}$

instance space




train
dev
test
$\mathcal{X}$
instance space

Experiment design

	training	development	testing
size	80%	10%	10%
purpose	training models	model selection	evaluation; never look at it until the very end

Binary classification

- Binary classification: $|y| = 2$
  [one out of 2 labels applies to a given x]

task	x	y
spam classification	email	{spam, not spam}

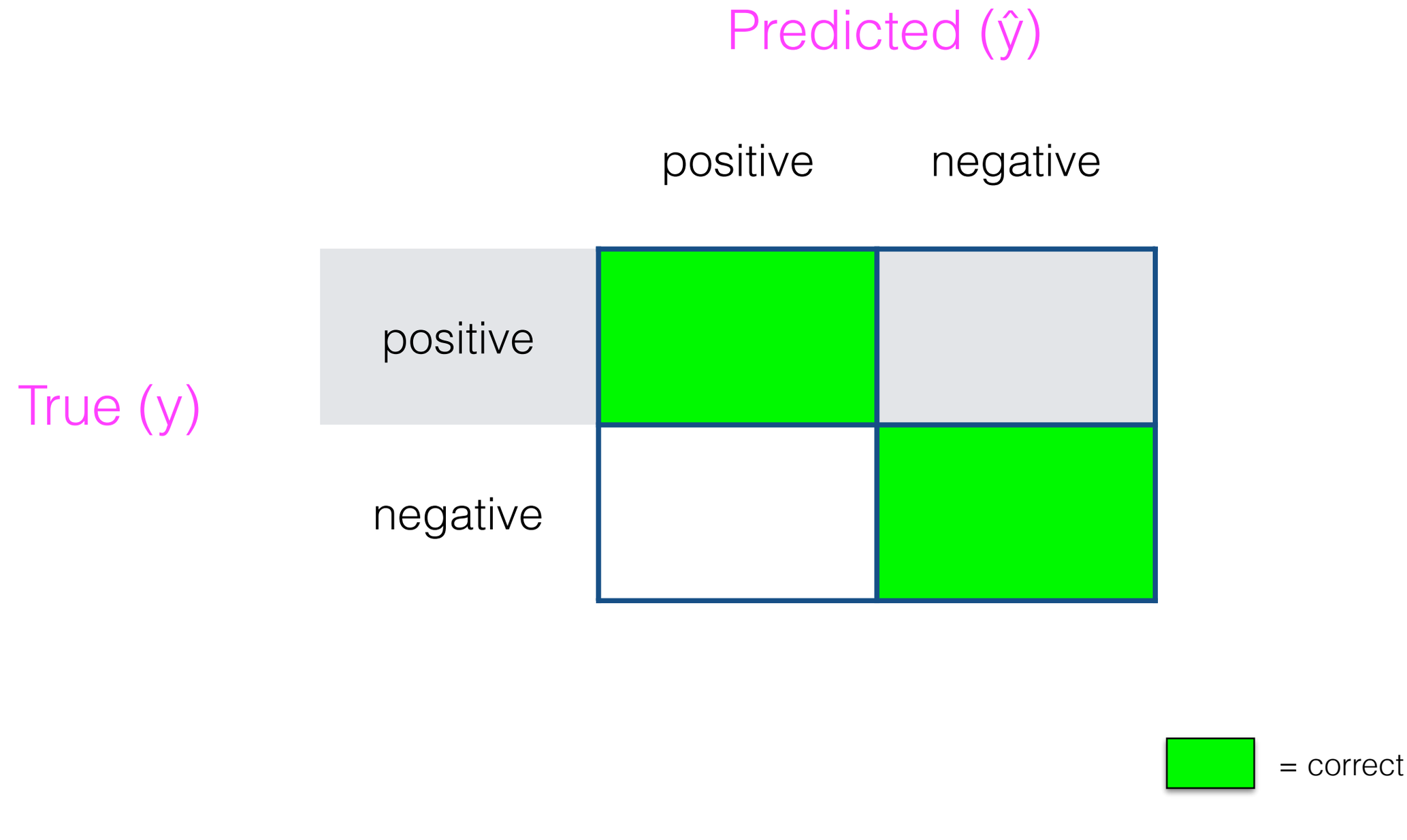
Accuracy

$$\text{accuracy} = \frac{\text{number correctly predicted}}{N}$$

$$\frac{1}{N}\sum_{i=1}^{N} I[\hat{y}_i = y_i] \qquad I[x] = \begin{cases} 1 & \text{if } x \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Perhaps most intuitive single statistic when the number of positive/negative instances are comparable

Confusion matrix



Confusion matrix

Accuracy = 99.3%

True (y) \ Predicted (ŷ)	positive	negative
positive	48	70
negative	0	10,347

= correct

Sensitivity

Sensitivity: proportion of true positives actually predicted to be positive

(e.g., sensitivity of mammograms = proportion of people with cancer they identify as having cancer)

a.k.a. “positive recall,” “true positive”

$$\frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i = \text{pos})}{\sum_{i=1}^{N} I(y_i = \text{pos})}$$

True (y) \ Predicted (ŷ)	positive	negative
positive	48	70
negative	0	10,347

Specificity

Specificity: proportion of true negatives actually predicted to be negative

(e.g., specificity of mammograms = proportion of people without cancer they identify as not having cancer)

a.k.a. “true negative”

$$\frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i = \text{neg})}{\sum_{i=1}^{N} I(y_i = \text{neg})}$$

True (y) \ Predicted (ŷ)	positive	negative
positive	48	70
negative	0	10,347

Precision

Precision: proportion of predicted class that are actually that class.
I.e., if a class prediction is made, should you trust it?

Precision(pos) =

$$\frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i = pos)}{\sum_{i=1}^{N} I(\hat{y}_i = pos)}$$

True (y) \ Predicted (ŷ)	positive	negative
positive	48	70
negative	0	10,347

Baselines

- No metric (accuracy, precision, sensitivity, etc.) is meaningful unless contextualized.
  - Random guessing/majority class (balanced classes = 50%, imbalanced can be much higher)
  - Simpler methods (e.g., election forecasting)

Scores

- Binary classification results in a categorical decision (+1/-1), but often through some intermediary score or probability

$$\hat{y} = \begin{cases} 1 & \text{if } \sum_{i=1}^{F} x_i \beta_i \geq 0 \\ -1 & 0 \text{ otherwise} \end{cases}$$

Perceptron decision rule

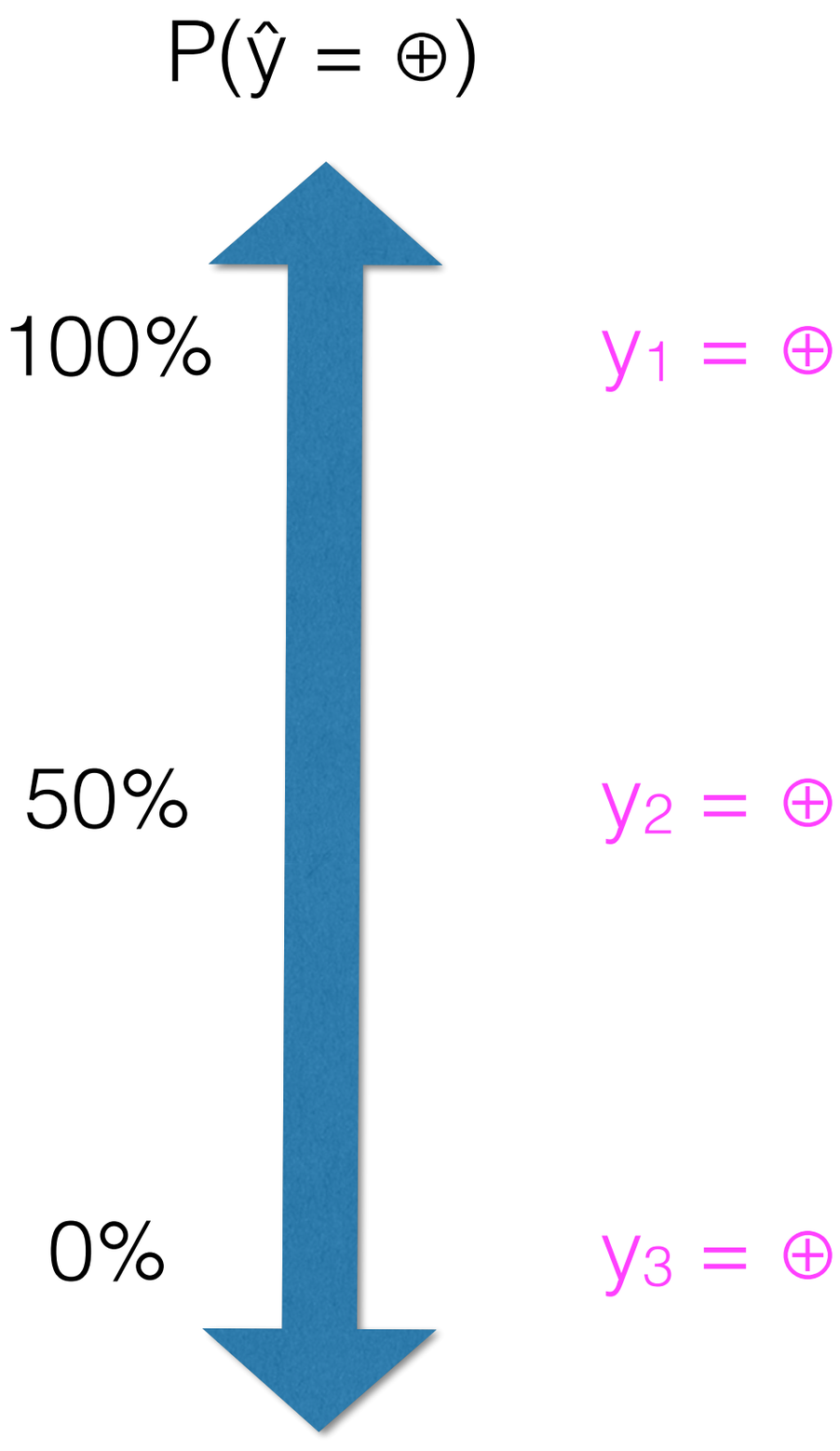
Scores

- The most intuitive scores are probabilities:

$$P(x = \text{pos}) = 0.74$$

$$P(x = \text{neg}) = 0.26$$

Instance Accuracy



Accuracy, precision, recall scores give a view of *model* accuracy, but we can also examine the predictions of individual data points

Multilabel Classification

- Multilabel classification: |y| > 1
  [multiple labels apply to a given x]

task	$\boldsymbol{x}$	y
image tagging	image	{fun, B&W, color, ocean, …}

Multilabel Classification

- For label space $\mathcal{Y}$, we can view this as $|\mathcal{Y}|$ binary classification problems
- Where y^j and y^k may be dependent
- (e.g., what's the relationship between y^2 and y^3?)

y	fun	0
y	B&W	0
y	color	1
y	sepia	0
y	ocean	1

Multiclass Classification

- Multiclass classification: $|y| > 2$

  [one out of N labels applies to a given x]

task	x	y
authorship attribution	text	{jk rowling, james joyce, …}
genre classification	song	{hip-hop, classical, pop, …}

Multiclass confusion matrix

True (y) \ Predicted (ŷ)	Democrat	Republican	Independent
Democrat	100	2	15
Republican	0	104	30
Independent	30	40	70

Precision

Precision(dem) =

$$\frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i = dem)}{\sum_{i=1}^{N} I(\hat{y}_i = dem)}$$

Precision: proportion of predicted class that are actually that class.

True (y) \ Predicted (ŷ)	Democrat	Republican	Independent
Democrat	100	2	15
Republican	0	104	30
Independent	30	40	70

Recall

$$\text{Recall(dem)} = \frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i = \textit{dem})}{\sum_{i=1}^{N} I(y_i = \textit{dem})}$$

Recall = generalized sensitivity (proportion of true class actually predicted to be that class)

True (y) \ Predicted (ŷ)	Democrat	Republican	Independent
Democrat	100	2	15
Republican	0	104	30
Independent	30	40	70

	Democrat	Republican	Independent
Precision	0.769	0.712	0.609
Recall	0.855	0.776	0.500

Predicted (ŷ)

True (y)

	Democrat	Republican	Independent
Democrat	100	2	15
Republican	0	104	30
Independent	30	40	70

Computational Social Science

- Lazer et al. (2009), Computational Social Science, Science.

- Grimmer (2015), We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together, APSA.

Computational Social Science

- Unprecedented amount of born-digital (and digitized) information about human behavior
    - voting records of politicians
    - online social network interactions
    - census data
    - expression of opinion (blogs, social media)
    - search queries

- Project ideas: “enhancing understanding of individuals and collectives”

Computational Social Science

- Draws on long traditions and rich methodologies in experimental design, sampling bias, causal inference. Accurate inference requires “thoughtful measurement”

- All methods have assumptions; part of scholarship is arguing where and when those assumptions are ok

- Science requires replicability. Assume your work will be replicated and document accordingly.