Week 3

Cogs 109: Data Analysis and Modeling

Fall 2017

Prof. Eran Mukamel

Regression with categorical predictors

- The linear regression framework can be applied using either quantitative or categorical predictors, or a combination of both
- To do this, we have to represent categorical predictors by numbers called dummy variables
- Why dummy? Because the numbers themselves don't mean anything
    - Example 1: Chocolate = 1, Vanilla = 0
    - Example 2: Chocolate = 0, Vanilla = 1
    - Example 3: Chocolate = -1, Vanilla = 1
- In all three cases, the math may look different but the results are equivalent

Categorical variable with 2 values (levels)

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

TABLE 3.7. *Least squares coefficient estimates associated with the regression of* `balance` *onto* `gender` *in the* `Credit` *data set. The linear model is given in (3.27). That is, gender is encoded as a dummy variable, as in (3.26).*

Alternative parameterization of the same data

First parameterization

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Second parameterization

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ -1 & \text{if } i\text{th person is male} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 - \beta_1 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

- The values of beta0, beta1 will be different
- However, the predictions (yhat) and statistics (p-value, t-value) will be the same

Categorical predictors with >2 values (levels)

- Need a separate dummy variable for each new level.
- For a factor with p levels, we would have p-1 dummy variables (plus the intercept)
- Example: Race = African-American/Caucasian/Asian

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases} \qquad x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American.} \end{cases}$$

Baseline

Extra effect of being Caucasian

Extra effect of being Asian

Regression results will have separate coefficients, p-values for each dummy variable

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American.} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`ethnicity[Asian]`	-18.69	65.02	-0.287	0.7740
`ethnicity[Caucasian]`	-12.50	56.68	-0.221	0.8260

Classification

MNIST database of handwritten digits

- http://yann.lecun.com/exdb/mnist/
- Humans easily and quickly recognize which category (0, 1, 2, 3, 4,… 9) a given image is in
- Many scientific and practical problems require classification

Classification vs. Regression

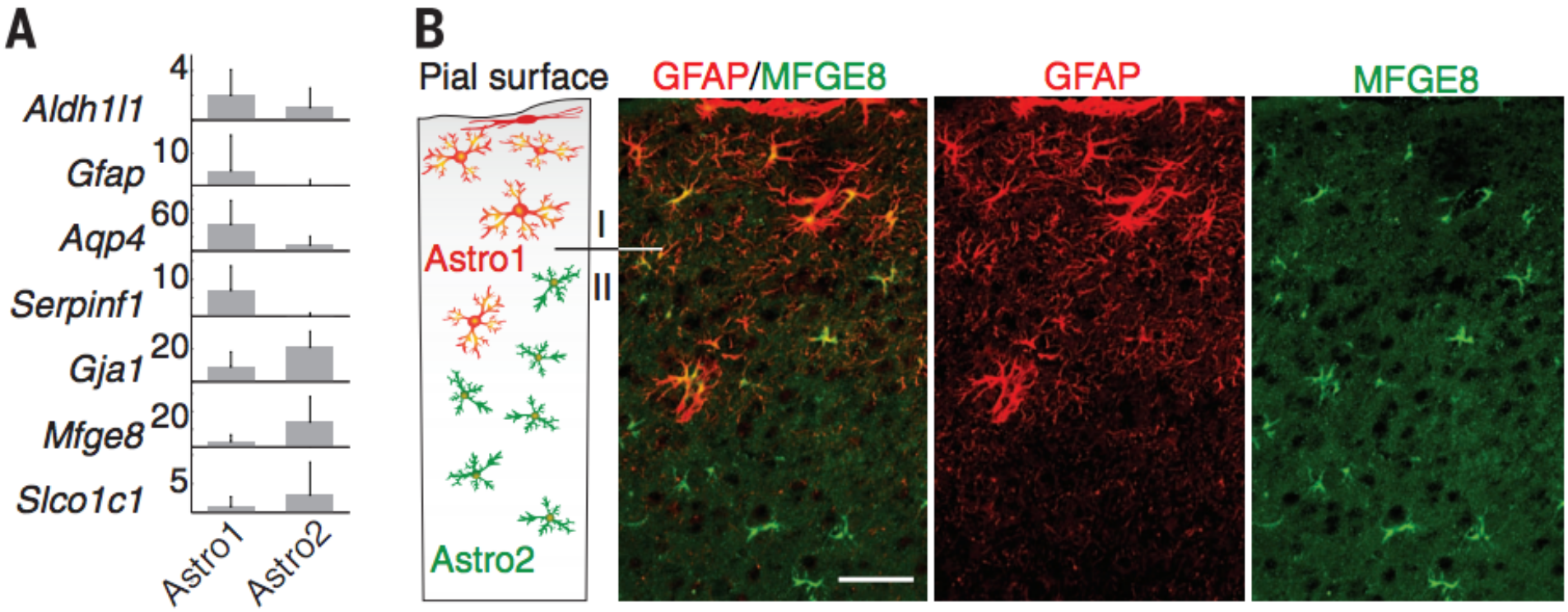
Regression

- Predictors: X
  - Quantitative, Qualitative/ Categorical, or a mix of both
- Outcome: y
  - Quantitative/Continous

Classification

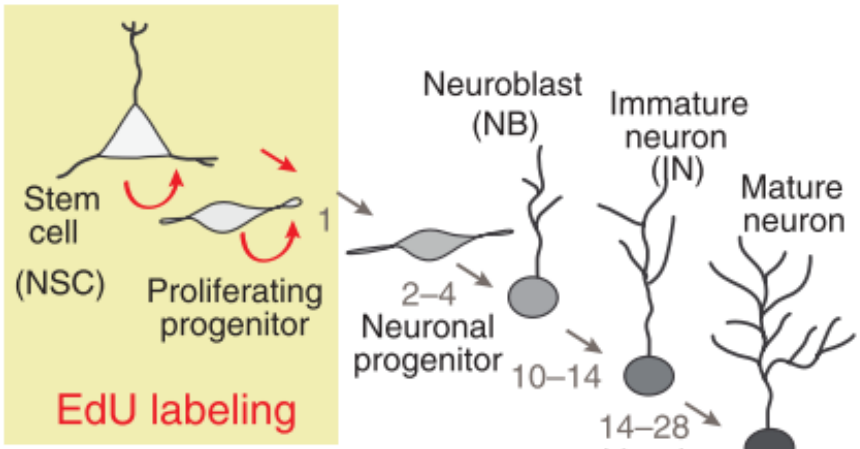
- Predictors: X
  - Quantitative, Qualitative/ Categorical, or a mix of both
- Outcome: y
  - Categorical/qualitative

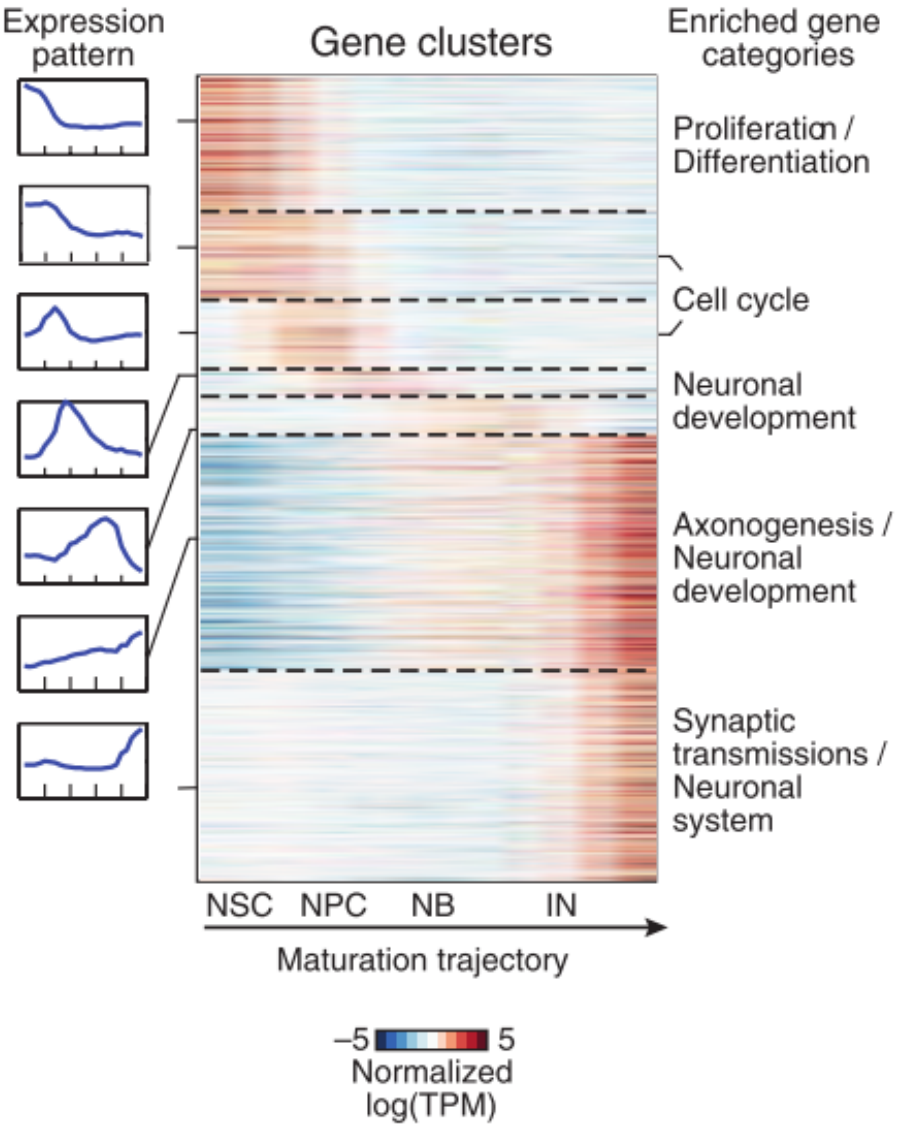
Classifying brain cell types by gene expression



A. Zeisel et al., Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142 (2015).

Classifying mature vs. immature neurons based on gene expression





http://birdcast.info/

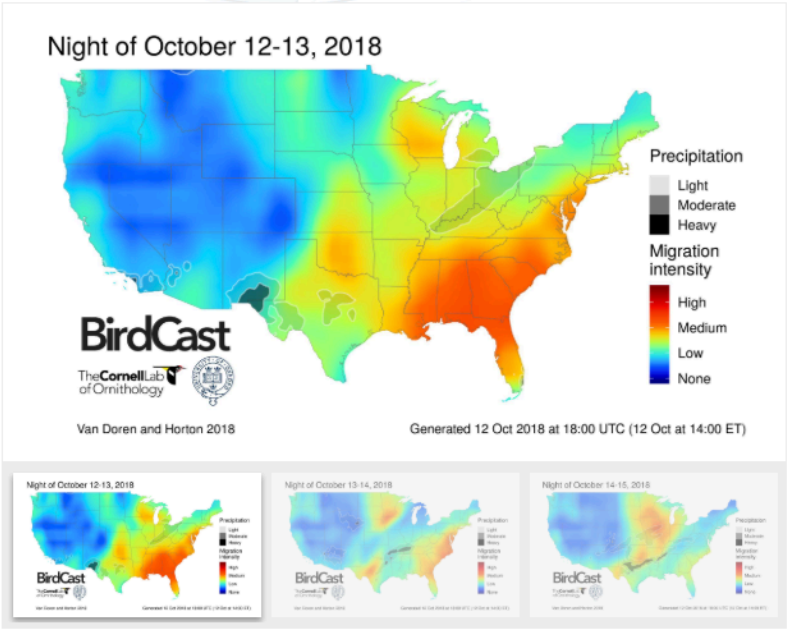
BirdCast

Research | People | Scientific Discussions | Live Migration Maps | Search for...

Bird Migration Forecasts in Real-Time

When, where, and how far will birds migrate? Our migration forecasts will answer these questions for the first time.

Migration Forecast



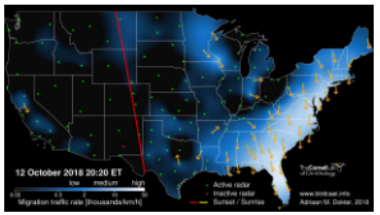
Bird migration forecasts powered by 23 years of radar observations and the most recent North American Mesoscale weather forecast. Migration forecasts show predicted nocturnal migration 3 hours after local sunset and are updated every 6 hours. Learn more

Scientific Discussion

Tracking Hurricane Michael

Hurricane Michael strengthened rapidly into a major hurricane on Tuesday, and as with previous storms on which BirdCast reported, it may have dramatic impacts on local and transient bird communities and their habitats when it comes ashore and passes through the Southeastern US. Live sightings will appear on the current observations map as they are entered into eBird, but as always, for those in the path of this storm, safety first!

Live Migration Maps



Real-time bird migration maps show intensities of actual bird migration as detected by the US weather surveillance radar network. All graphics are relative to the Eastern time zone.

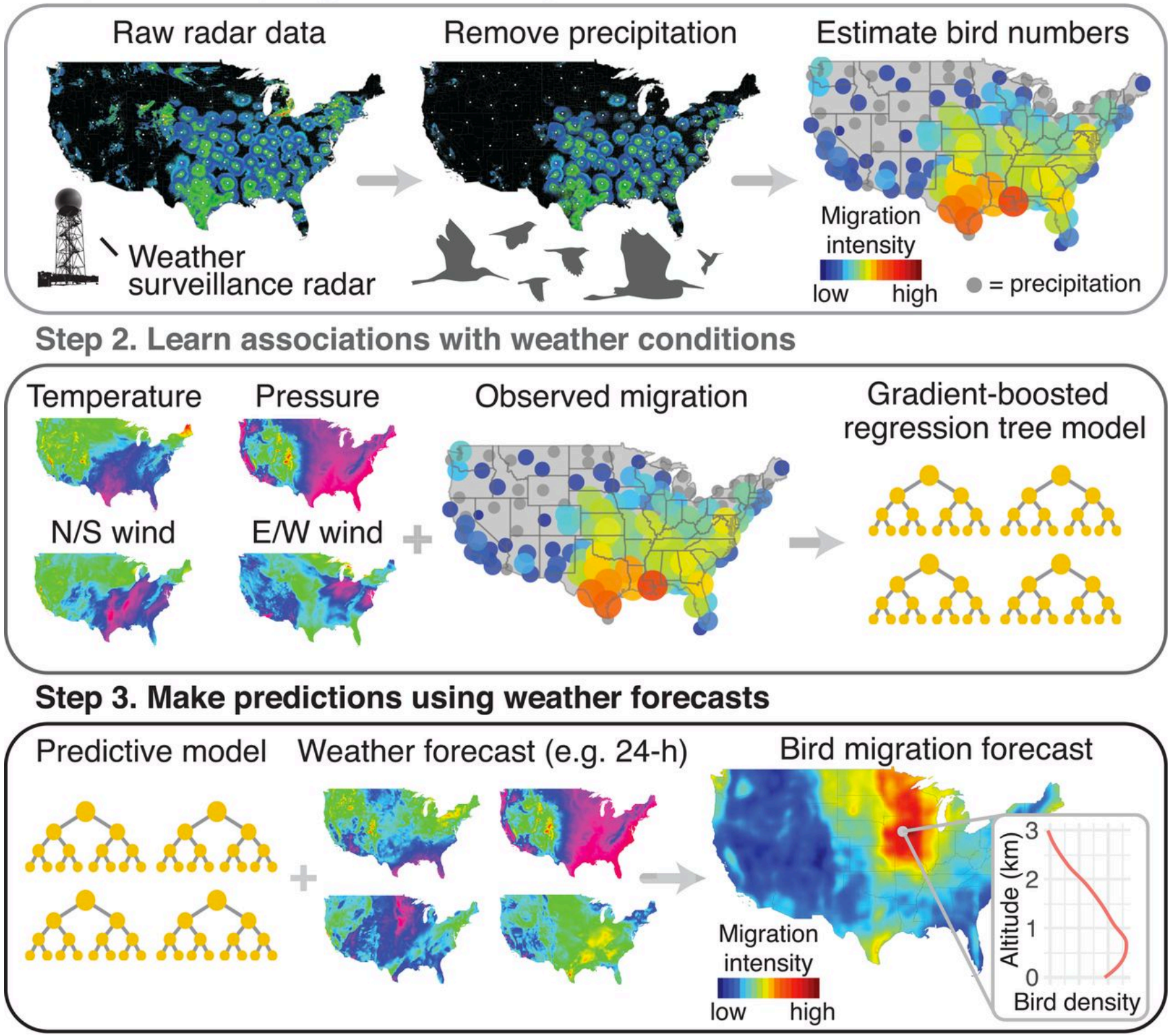
View Live Migration Maps

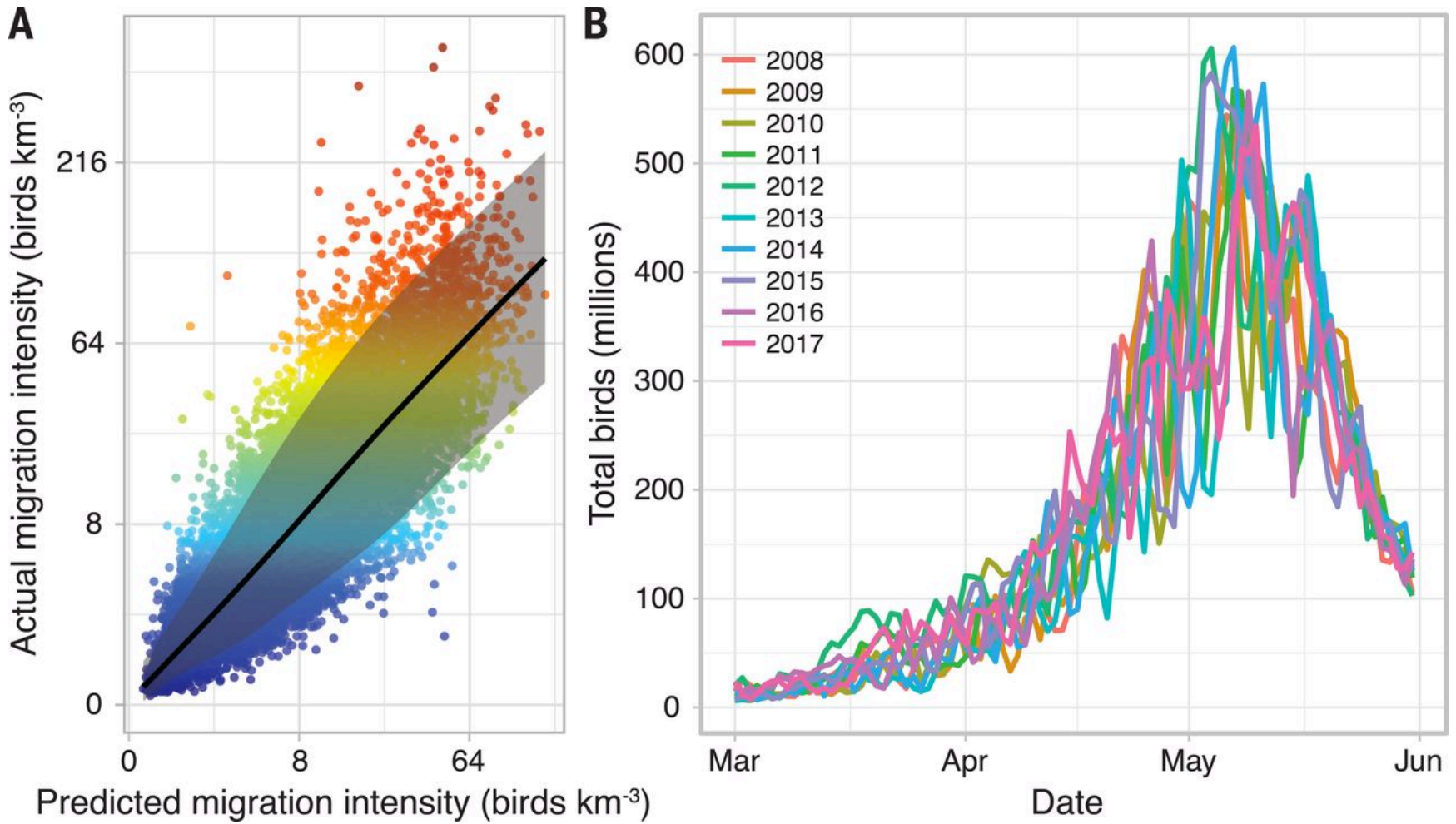
Tweets by @DrBirdCast

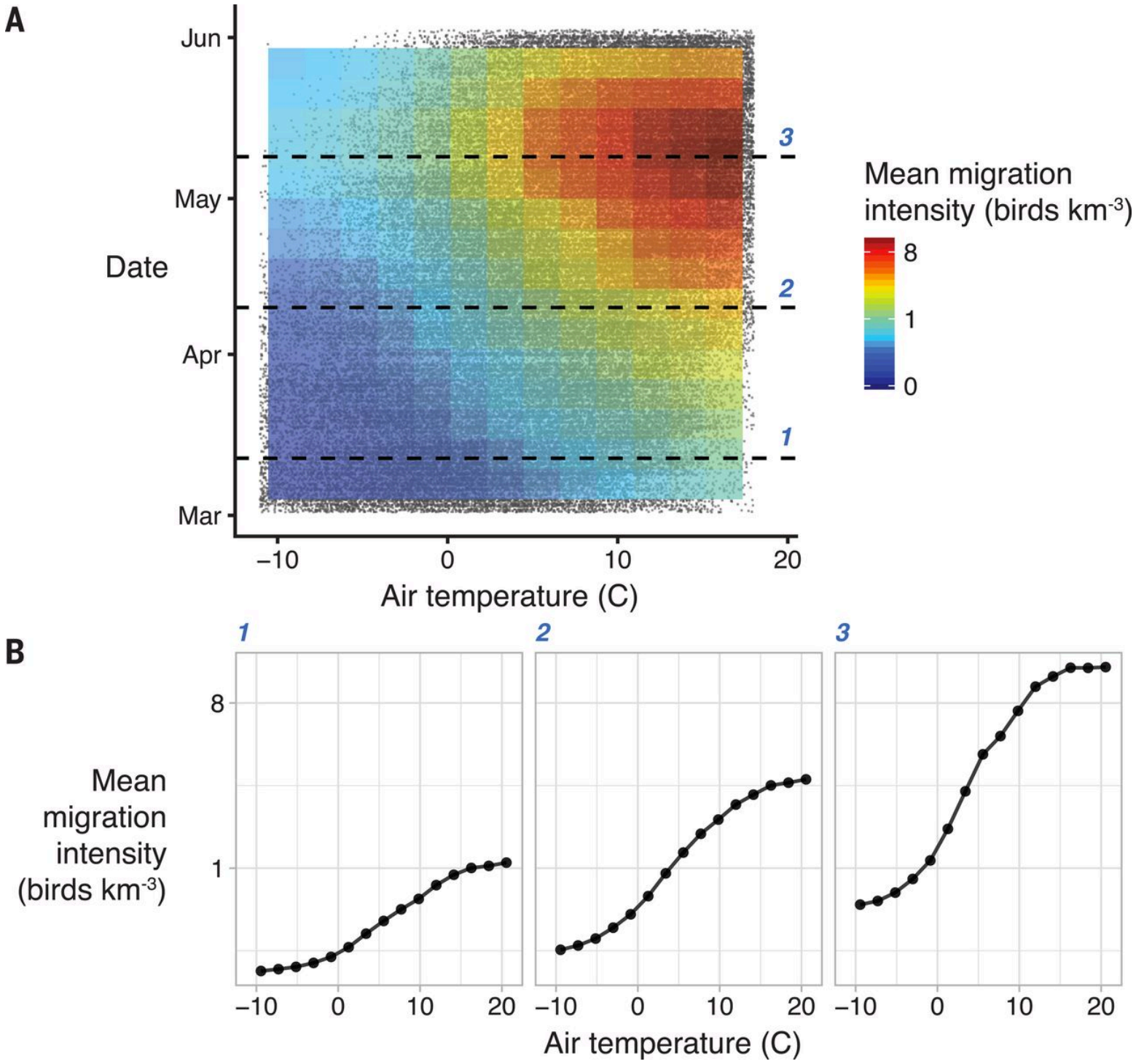
BirdCast–Cornell Lab Retweeted

Houston Audubon
@HoustonAudubon

🚨LIGHTS OUT ACTION ALERT 🚨 Houston Audubon recommends buildings three stories and higher turn LIGHTS OUT tonight and tomorrow night (Oct. 10-13). Conditions are favorable for intense bird migration while a frontal system is moving through the region.

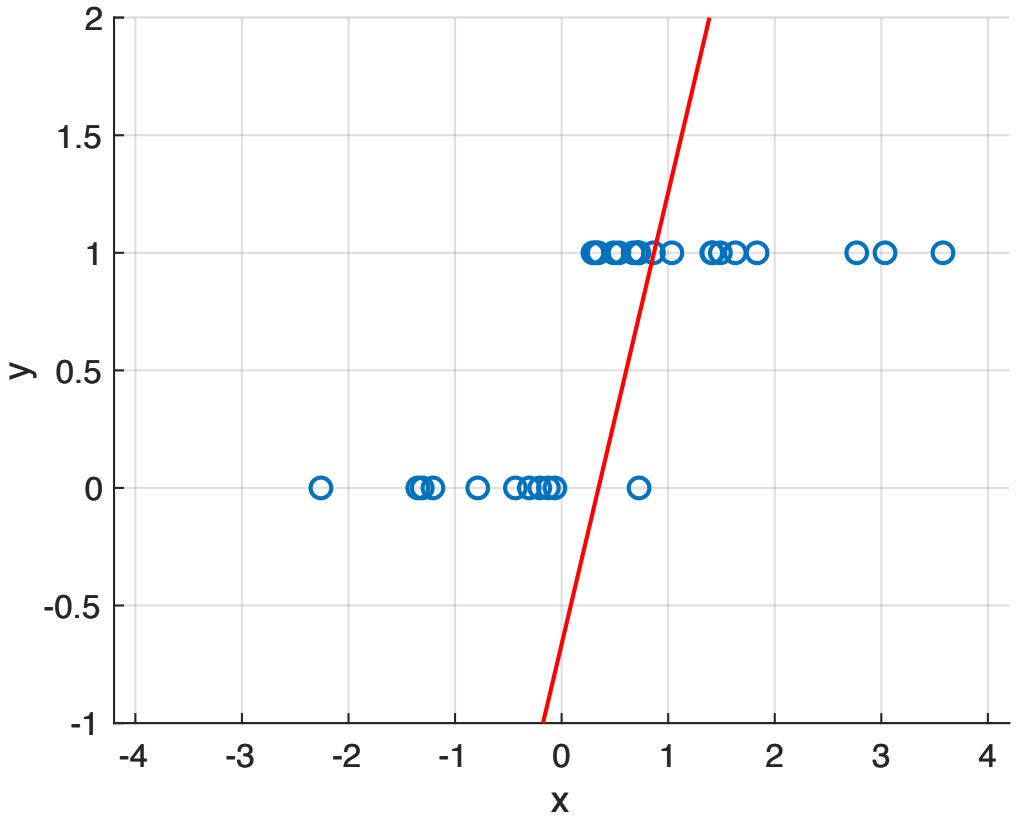

Step 1. Quantify migration intensity at 143 weather radar stations
Raw radar data
Remove precipitation
Estimate bird numbers
Weather surveillance radar
Migration intensity
low
high
= precipitation
Step 2. Learn associations with weather conditions
Temperature
Pressure
Observed migration
Gradient-boosted regression tree model
N/S wind
E/W wind
Step 3. Make predictions using weather forecasts
Predictive model
Weather forecast (e.g. 24-h)
Bird migration forecast
Migration intensity
low
high
Altitude (km)
3
2
1
0
Bird density


A
Actual migration intensity (birds km-3)
216
64
8
0
0
8
64
Predicted migration intensity (birds km-3)
B
Total birds (millions)
600
500
400
300
200
100
0
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Mar
Apr
May
Jun
Date


A
Jun
May
Apr
Mar
Date
3
2
1
−10
0
10
20
Air temperature (C)
Mean migration intensity (birds km-3)
8
1
0
B
1
2
3
8
1
Mean migration intensity (birds km-3)
−10
0
10
20
Air temperature (C)

Why not use linear regression?

Reason #1: Linear regression gives nonsensical values outside the bounds of our categorical variable



Why not use linear regression?

Reason #2: Categorical values may not necessarily be arranged in a unique order

Example: X = {Age, Gender, Time of day, etc.}, Y = {chocolate, vanilla, strawberry}

We could try to “encode” Y as a numerical value, e.g.
Chocolate=0, Vanilla=1, Strawberry=2
Or: Chocolate=0, Strawberry=1, Vanilla=2
Or: Chocolate=0, Vanilla=1, Strawberry=20

Each of these would give a different result!

Instead, we would like a model that treats Chocolate/Vanilla/Strawberry as separate categories

Logistic regression

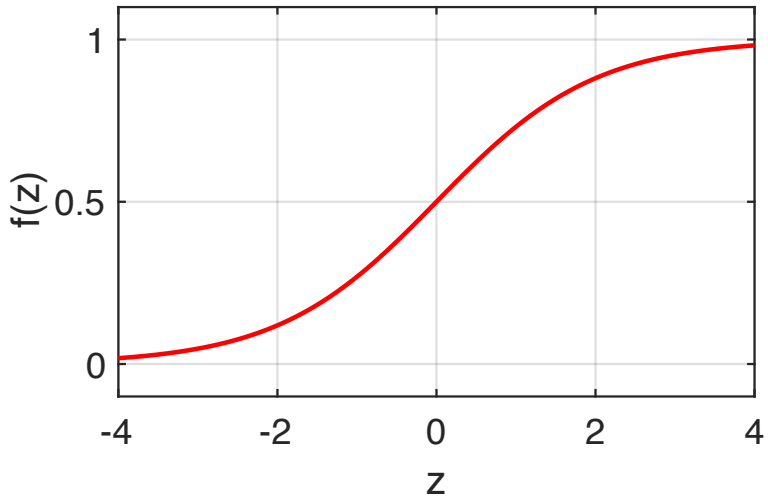
- This is a useful extension of linear regression when y has 2 classes (e.g. y=0 or 1, or y=Male or Female)
- Goal: Predict the probability that y=1, P(y=1|x)
- Note that y=0 or 1, but P(y) is a continuous value in the range [0,1]

Linear regression: $y = \beta_0 + \beta_1 X + \varepsilon$

Logistic regression: $Prob(y = 1|x) = f(\beta_0 + \beta_1 X)$

"Logistic" function $f(z) = e^z/(1 + e^z)$

$Prob(y = 0|x) = 1 - Prob(y = 1|x) = 1 - f(\beta_0 + \beta_1 X)$



Where did the logistic function come from?

- Instead of describing P(X), it is more convenient to talk about the *odds* (or odds ratio):

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}.$$

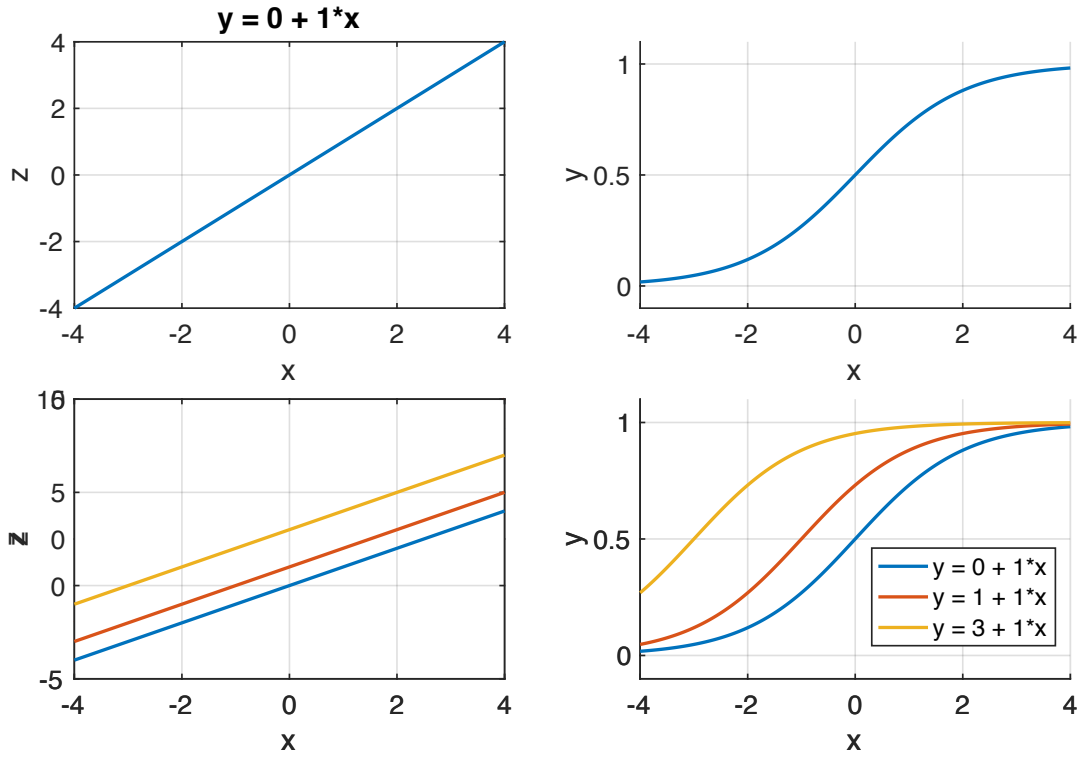
- Logistic regression assumes that the *log odds* (or *logit*) is linear in X:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X. \quad \longleftrightarrow \quad p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

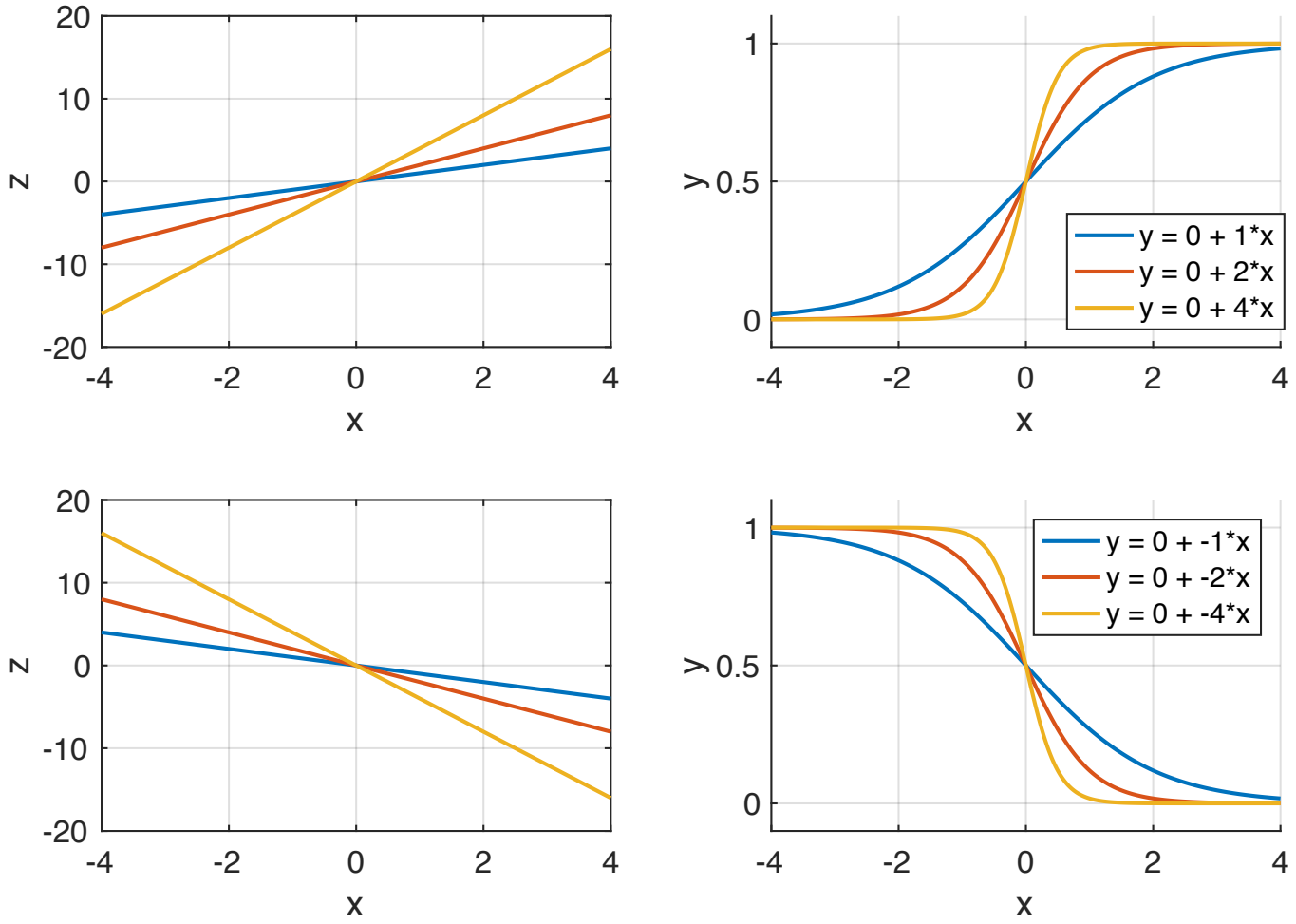
Intercept (beta0) shifts the prediction left and right

The logistic function is an “S-shaped” or “sigmoid” curve, which always gives a number between 0 and 1.

When X = -beta0/beta1, P(y|X) = 0.5 = 50%



Slope (beta1) determines steepness

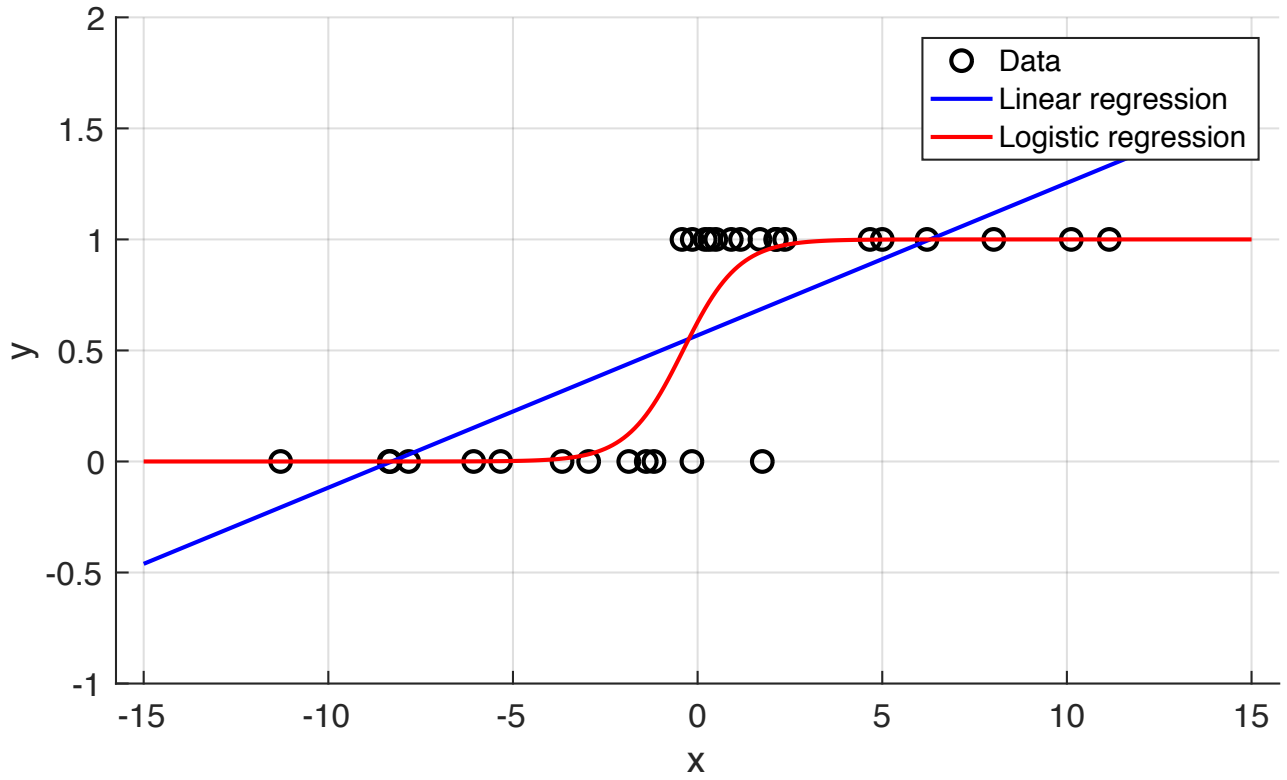


Fitting the logistic regression

- Want to find values of beta0 (intercept) and beta1 (slope) that give the “best” fit

- We could use least squares (as in linear regression), but it is better (and more common) to use maximum likelihood:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$



Generalized linear regression model:
    logit(y) ~ 1 + x1
    Distribution = Binomial

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	0.52382	0.66481	0.78792	0.43074
x1	1.3269	0.61483	2.1581	0.030919

30 observations, 28 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 26.8, p-value = 2.3e-07

Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	0.56816	0.064548	8.8022	1.487e-09
x1	0.068607	0.012334	5.5622	5.9814e-06

Number of observations: 30, Error degrees of freedom: 28
Root Mean Squared Error: 0.354
R-squared: 0.525, Adjusted R-Squared 0.508
F-statistic vs. constant model: 30.9, p-value = 5.98e-06

Multiple logistic regression

When we have >1 predictors, we can model the log-odds using a multiple linear regression:

Intercept

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

Slope for X1 Slope for Xp

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

Example:

Fast Lightning is a racehorse who normally has 1-to-1 odds of winning a race on a clear day. However, if it's cloudy outside then Fast Lightning's odds improve to 2-to-1. If it rains then Fast Lightning is favored at 5-to-1 odds.

Questions:

1. How can we represent Fast Lightning's chances using a logistic regression?
2. What is the value of beta0?
3. How many slope terms will the model have? What are the values of the slopes?
4. What is the probability that FL will win if it rains?

Multiple logistic regression

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

Example:

Fast Lightning is a racehorse who normally has 1-to-1 odds of winning a race on a clear day. However, if it's cloudy outside then Fast Lightning's odds improve to 2-to-1. If it rains then Fast Lightning is favored at 5-to-1 odds.

Questions:

1. How can we represent Fast Lightning's chances using a logistic regression?

2.What is the value of beta0?

3.How many slope terms will the model have? What are the values of the slopes?

4.What is the probability that FL will win if it rains?

Fast Lightning is a racehorse who normally has 1-to-1 odds of winning a race on a clear day. However, if it's cloudy outside then Fast Lightning's odds improve to 2-to-1. If it rains then Fast Lightning is favored at 5-to-1 odds.

Questions:

1. How can we represent Fast Lightning's chances using a logistic regression?
2. What is the value of beta0?
3. How many slope terms will the model have?
   What are the values of the slopes?
4. What is the probability that FL will win if it rains?

$$y = \{win, loss\}$$

Dummy variables:

$X_{cloudy} = 1$ for cloudy, 0 otherwise

$X_{rain} = 1$ for rain, 0 otherwise

$$\log \frac{P(y=win)}{P(y=loss)} = \beta_0 + \beta_{cloudy} X_{cloudy} + \beta_{rain} X_{rain}$$

If it's clear:

$$\log(P(y = win)/P(y = loss)) = \log(1) = 0$$
$$= \beta_0 + \beta_{cloudy} 0 + \beta_{rain} 0 = \beta_0$$

Thus: $\beta_0 = 0$

If it's cloudy:

$$\log(P(y = win)/P(y = loss)) = \log(2/1)$$
$$= \beta_0 + \beta_{cloudy} 1 + \beta_{rain} 0 = \beta_{cloudy}$$

Thus: $\beta_{cloudy} = \log(2)$

If it's rainy:

$$\log(P(y = win)/P(y = loss)) = \log(5/1)$$
$$= \beta_0 + \beta_{cloudy} 0 + \beta_{rain} 1 = \beta_{rain}$$

Thus: $\beta_{rain} = \log(5)$

$$P(y = win|rain) = exp(\beta_{rain})/(1 + \exp(\beta_{rain})) = 5/(1 + 5) = 5/6$$

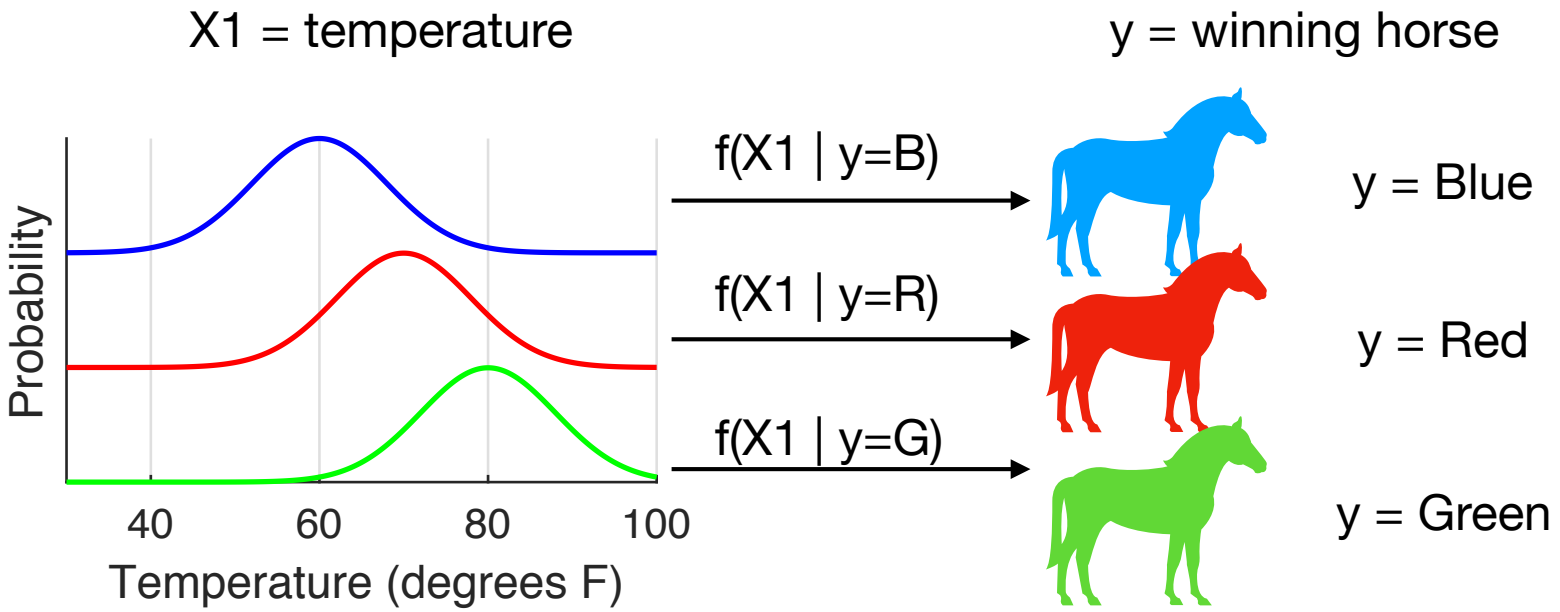
Classification for 3 or more classes: *Linear discriminant analysis (LDA)*

- Logistic regression is usually applied to binary classification (i.e. discriminating between 2 classes)
- When we have >2 classes, we can build a classifier by using *Bayes' theorem*.
- Example: Three racehorses (Blue, Red, Green) are competing, only one can win. Red is usually the fastest. However, when it is colder than usual Blue tends to win, while on hot days Green tends to win.

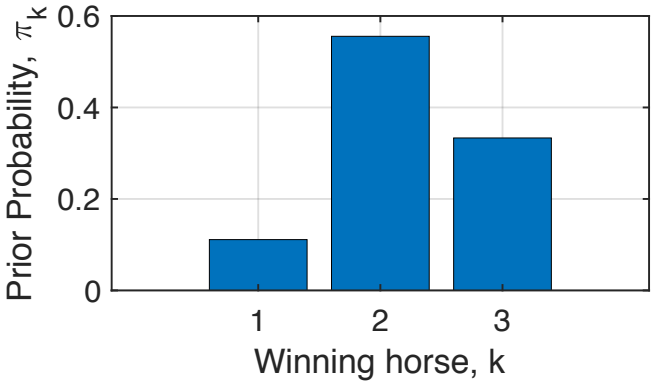
Classification for 3 or more classes: *Linear discriminant analysis (LDA)*

Start by describing (modeling) the probability of X conditioned on each class of y, f(X|y):

Which horse will win the race?



We also need to describe the *prior probability* that y=k, given that we don't know the temperature. For example, this could be



LDA

Bayes' theorem tells us that:

$$P(y \mid X) = \frac{P(X|y)P(y)}{\sum_y P(X|y)\,P(y)}$$

Prior

Posterior

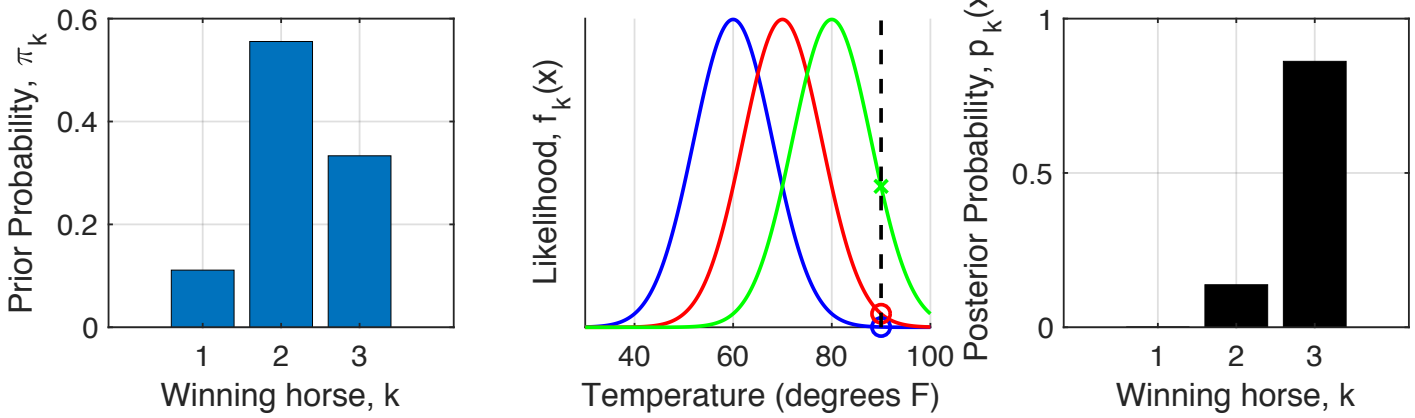
Data likelihood

Our book uses a different notation for prior (π) and likelihood (f):

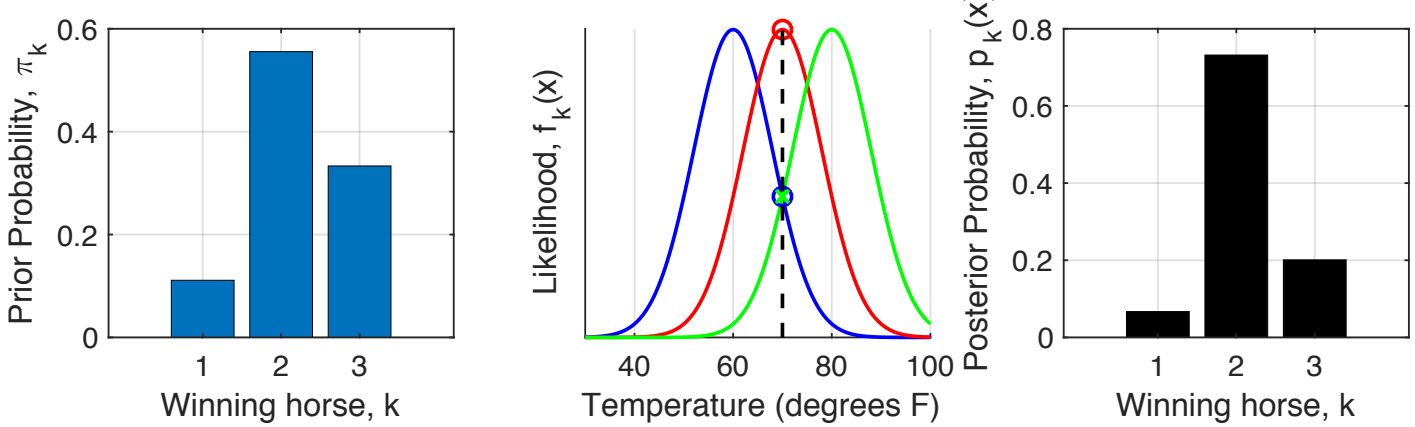
$$P(y = k \mid X = x) = p_k(x) = \frac{f_k(x)\pi_k}{\sum_y f_k(x)\pi_k}$$

$$p_k(x) = \frac{f_k(x)\pi_k}{\sum_y f_k(x)\pi_k}$$

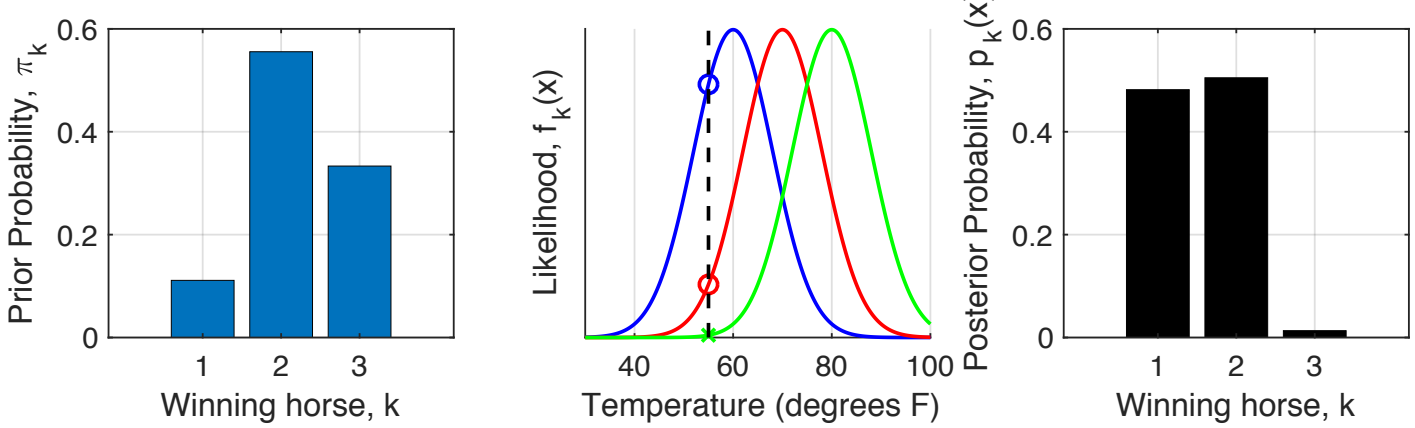
Hot day,
T=90



Mild day
T=70



Cold day,
T=50



LDA

- Assume the likelihood function is normal (Gaussian):

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right),$$

- Also assume that all the variances are equal, $\sigma_k = \sigma$ for all k

- Then the log posterior (i.e. log[$p_k(x)$]) is a simple linear function of x:

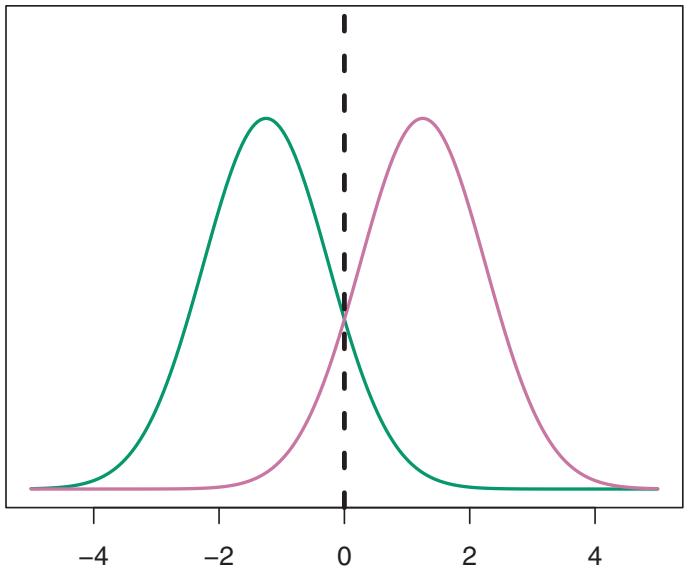
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

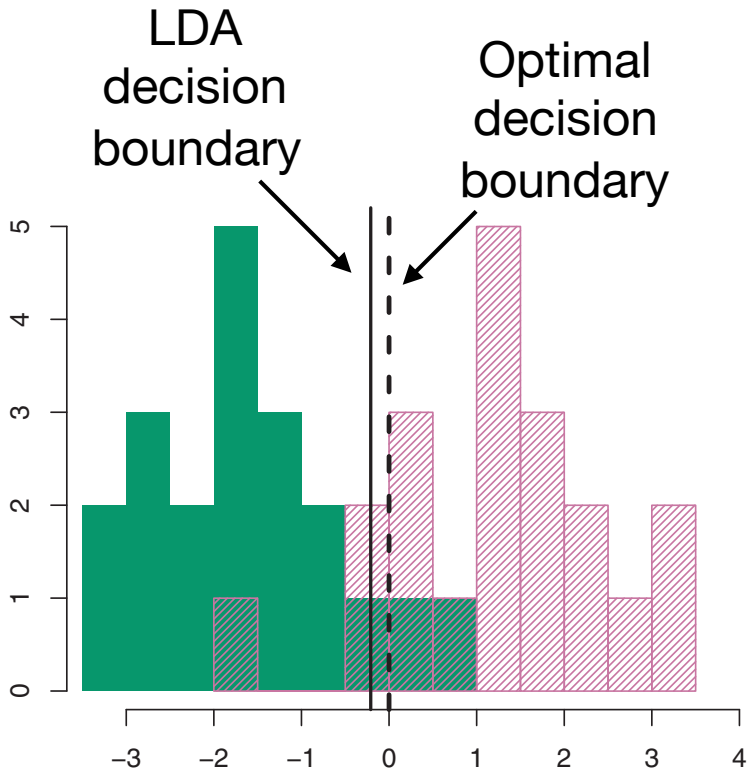
- The best (Bayesian, optimal) classifier will choose the class with highest posterior probability:

$$\hat{k} = \arg\max_k (x) = \arg\max_k \delta_k(x)$$

Example: 2 classes

Decision boundary





In practice, we don't know the values of μ and sigma, so we must estimate them from our data:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

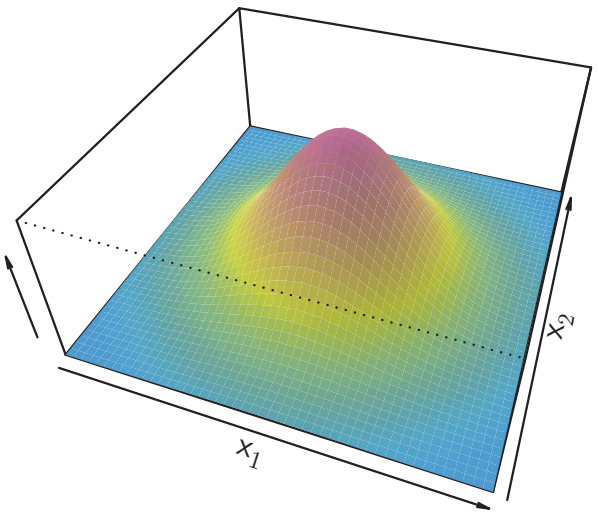
$$\hat{\pi}_k = n_k/n.$$

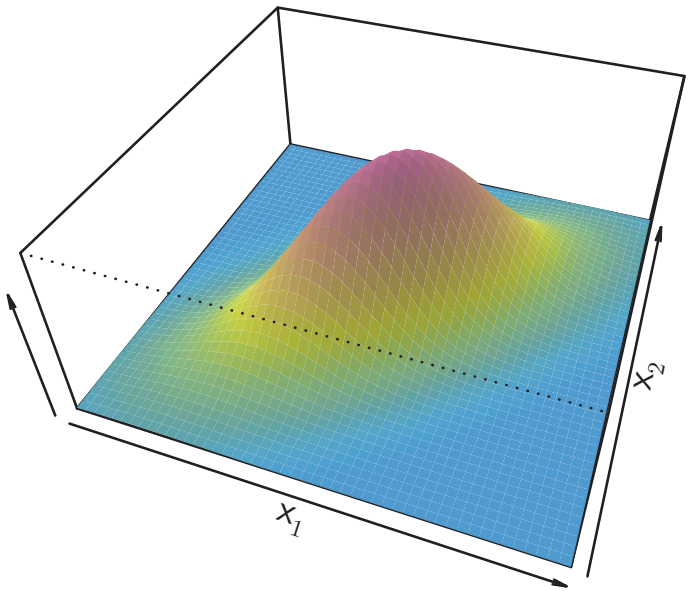
LDA with 2 or more predictors

- Assume that the predictors come from a multivariate normal (Gaussian) distribution:

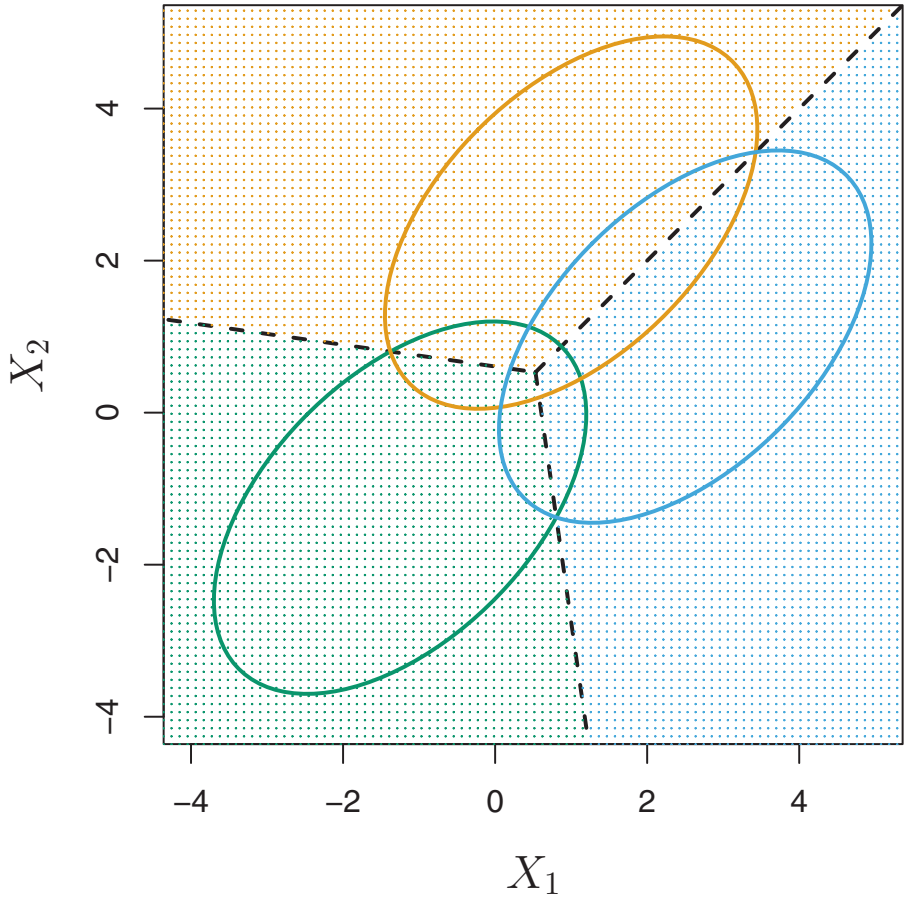
$$X = (X_1, X_2, \ldots, X_p)$$

$$f(x) \propto e^{-[a\,x_1{}^2 + b x_2{}^2 + \ldots + c x_1 x_2 + \ldots]}$$

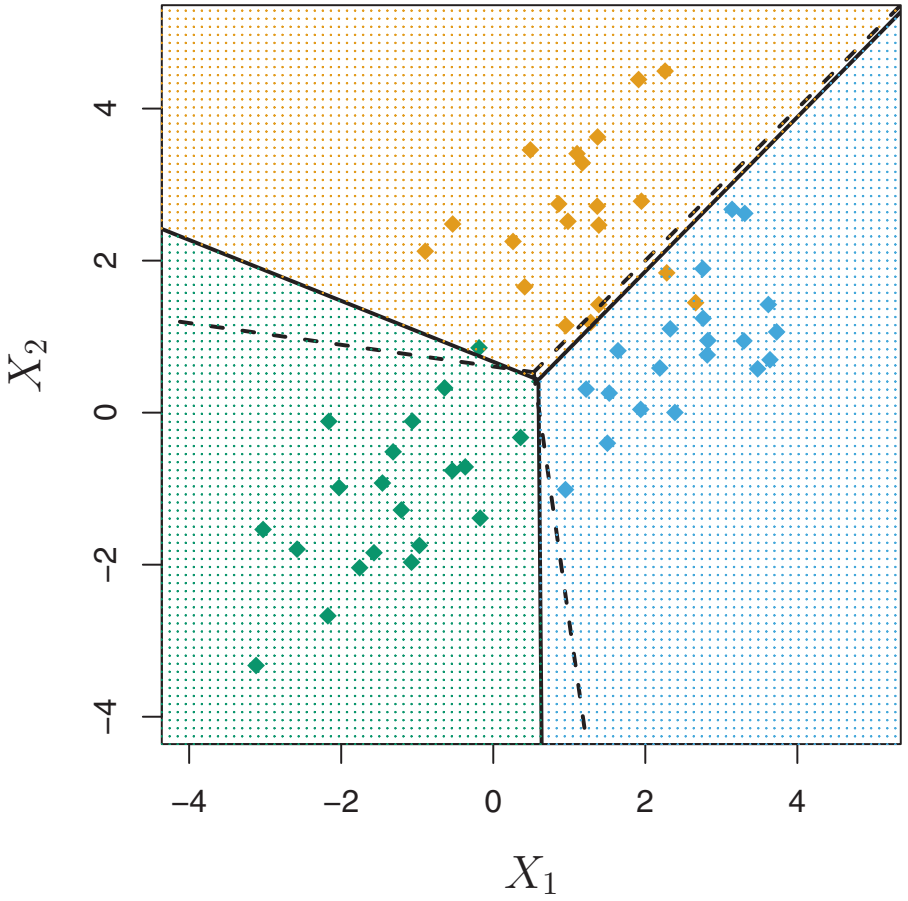




True data distributions (3 classes, each one producing a multivariate Gaussian for X1,X2) and optimal decision boundaries (dashed lines)



LDA classifier boundaries (solid lines)

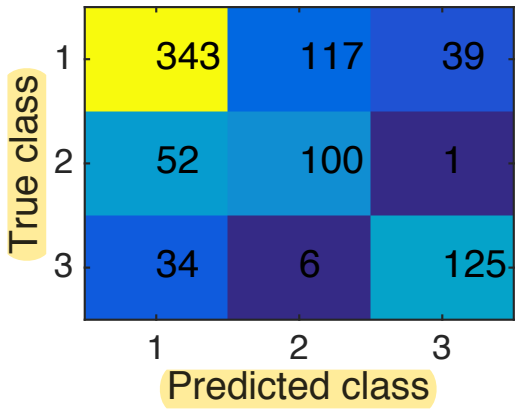


Evaluating classification results:

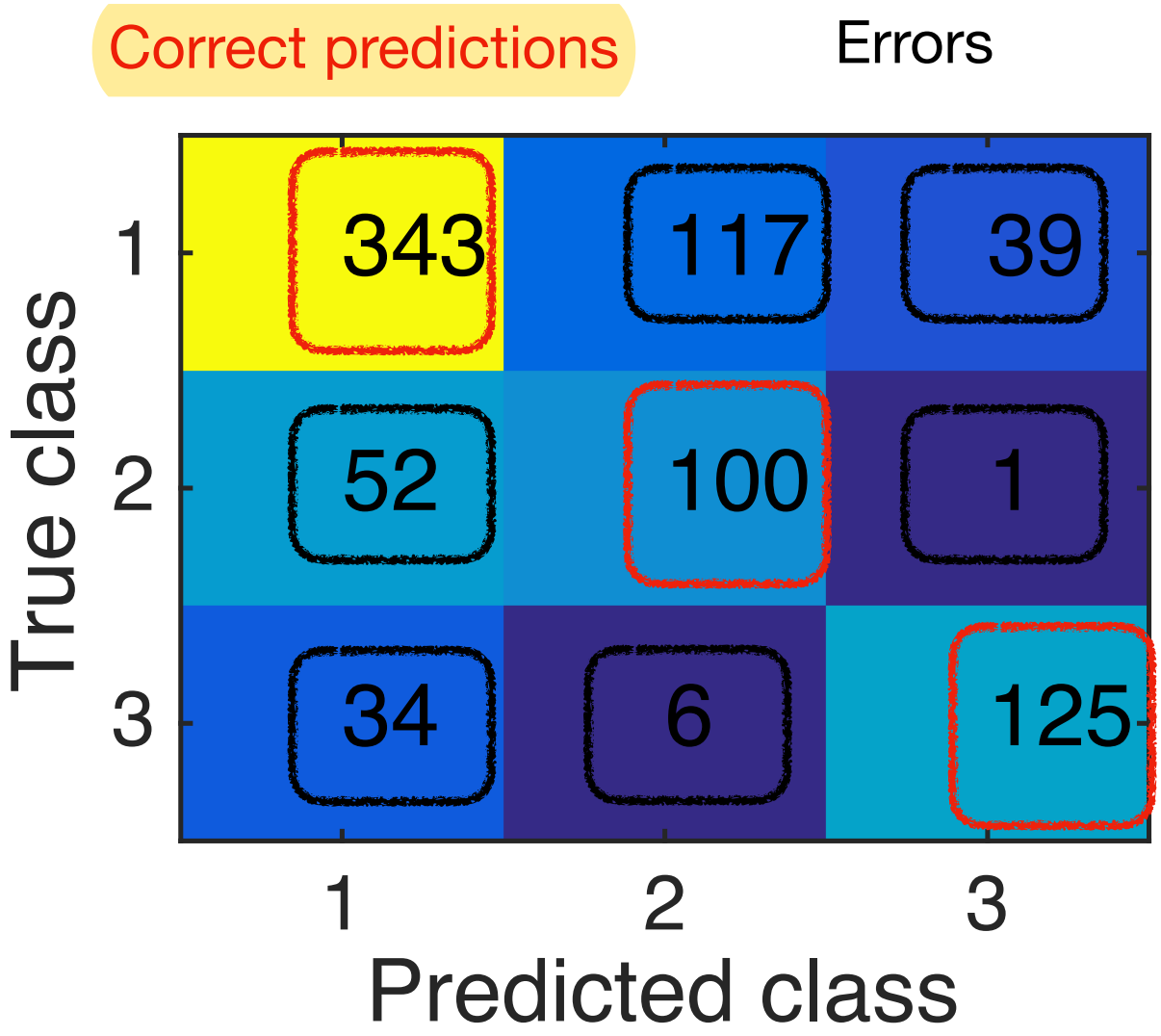
Confusion matrix

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

- Confusion matrix:
  - For every category i, how many observations were classified as j?



Confusion matrix



Comparing classifiers: Sensitivity and Specificity

- "Default" data: In this data set, most subjects (97%) did not default
- The LDA classifier has high *specificity*: % of non-defaulters that are correctly identified (9644/9667)
- However, it has poor *sensitivity*: % of defaulters who are correctly identified (81/333)
- We can modify LDA by lowering the decision threshold to capture more of the defaulters

LDA: Choose the most likely category; achieves lowest total error

$$\Pr(\texttt{default} = \texttt{Yes}|X = x) > 0.5.$$

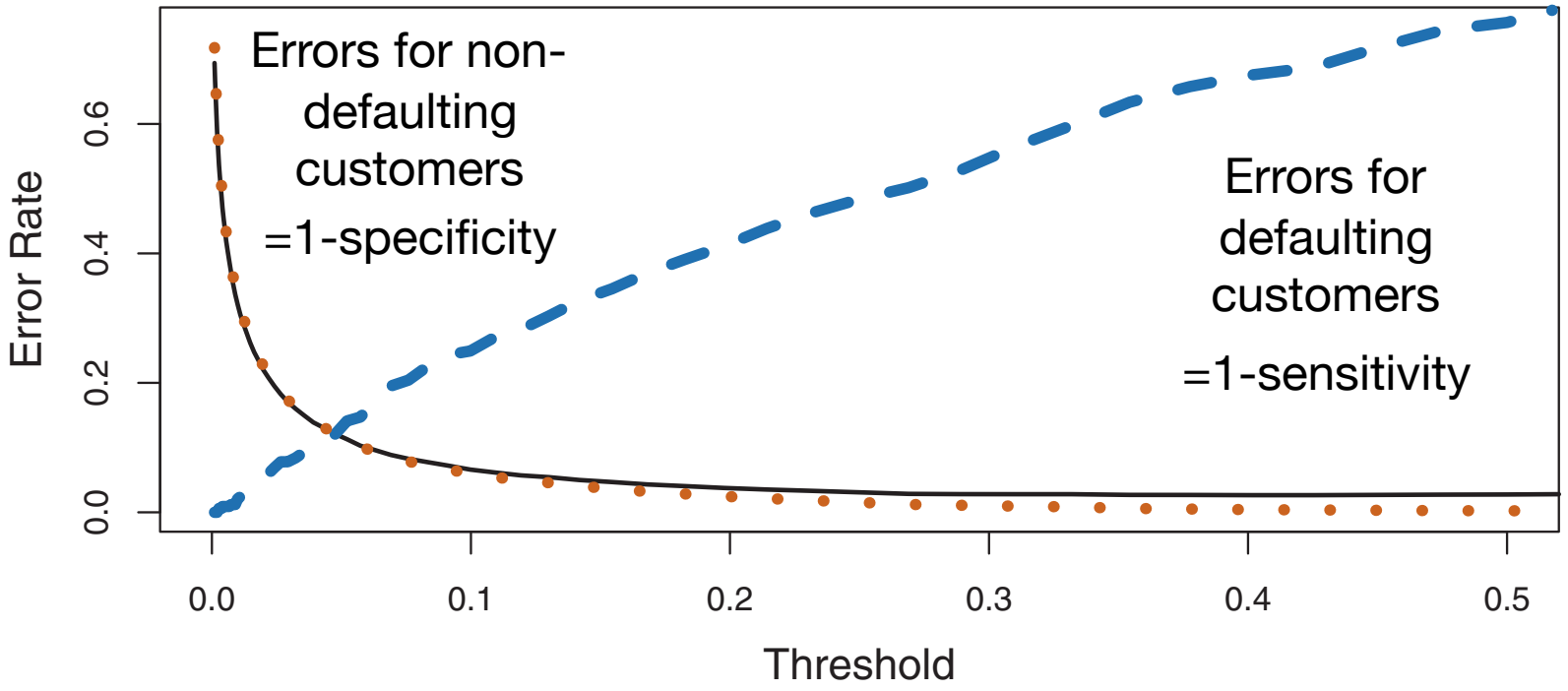
		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

LDA with modified threshold: does a better job on the subjects who default

$$P(\texttt{default} = \texttt{Yes}|X = x) > 0.2.$$

		True default status		
		No	Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

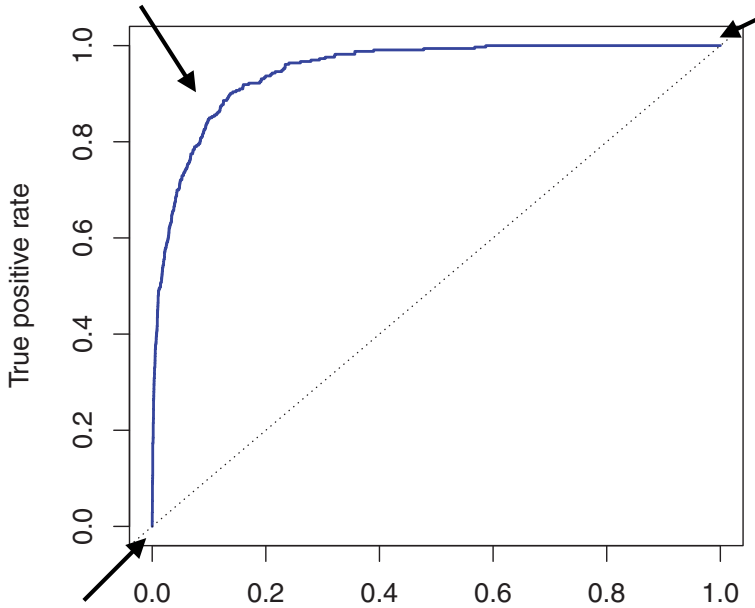
By varying the decision threshold, we can adjust the tradeoff between sensitivity and specificity



Receiver operating characteristic (ROC) curve shows the tradeoff

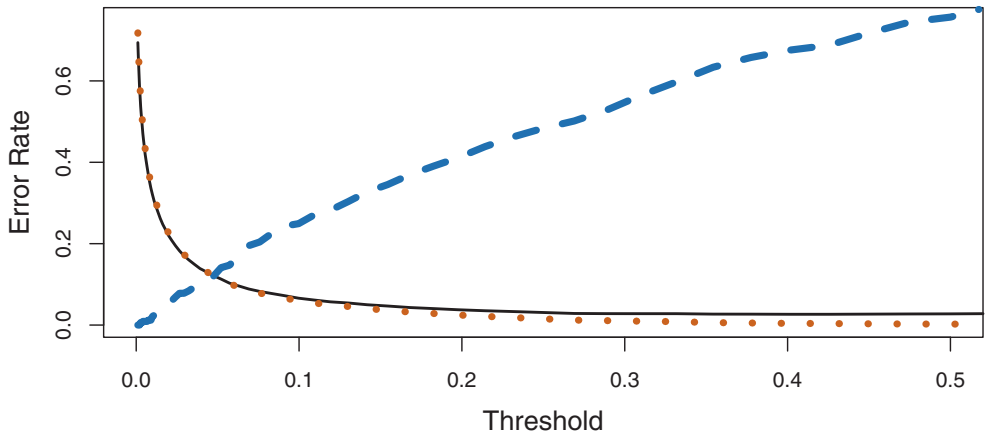
Good threshold choices: low FP and low FN

Low threshold, no false negatives

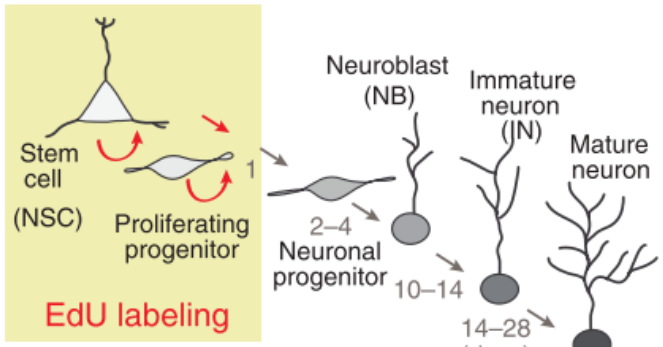


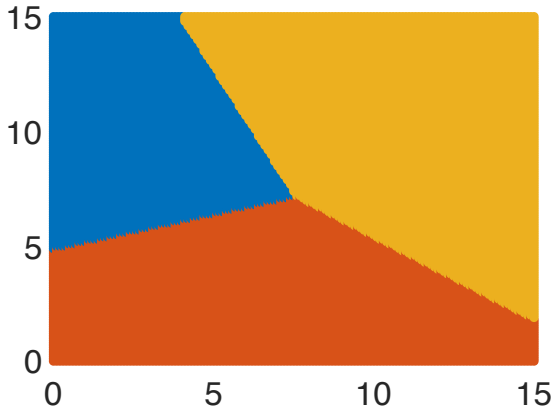
High threshold, no false positives

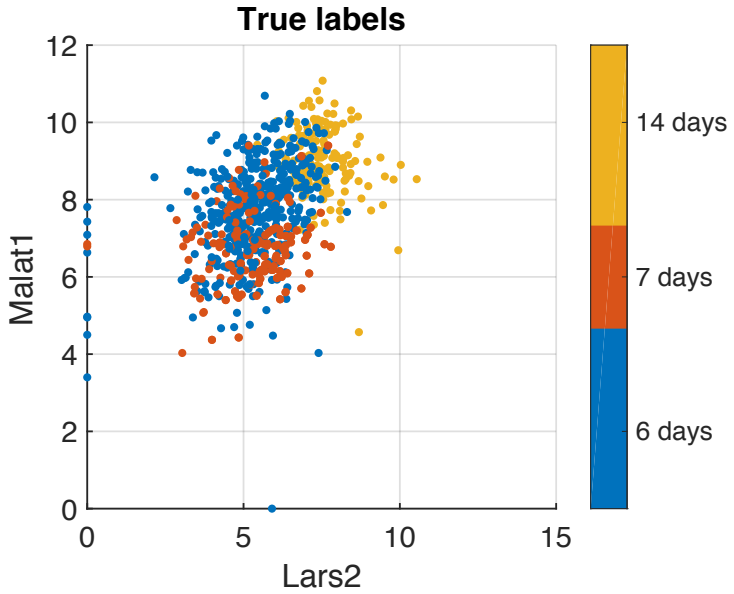
		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N^*	P^*	

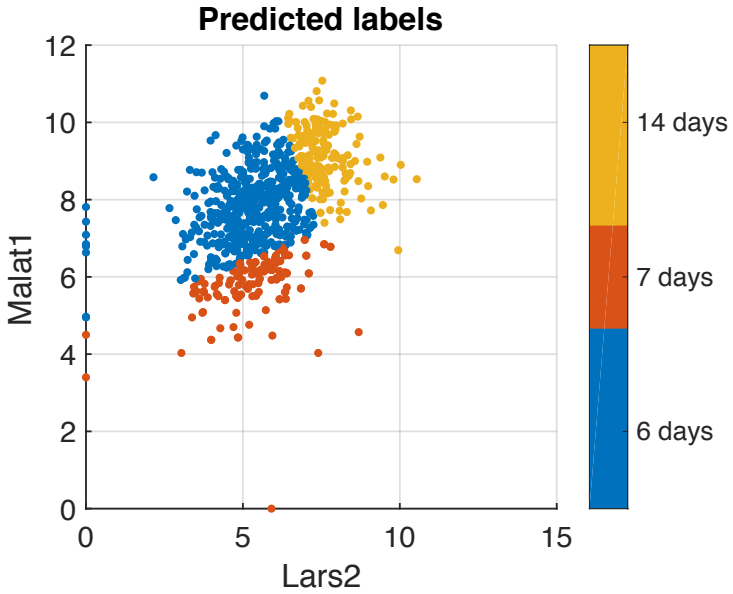


Example: Classifying neurons by birthdate









LDA vs. K-nearest neighbors

- Which one is parametric and which non-parametric?
- Which one is more flexible: LDA or KNN with k=3?
- Which would you expect to have more bias? More variance?