Data Science Bootcamp, 10th January 2017

Introduction to Machine Learning

James McInerney

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Key Moments for Machine Learning

- Dartmouth 1956 conference on artificial intelligence.
- Machine learning “gives computers the ability to learn without being explicitly programmed” (Arthur Samuel, 1959).
- “AI Winters” of 1970s and 1980s.
- Markov chain Monte Carlo 1990s.
- 2000s big data, faster computation, deep models.

Types of Machine Learning

Type	Observations	Gist	Examples
Supervised	inputs and outputs	learn function from input to output; maximize predictive accuracy	logistic regression, neural networks, K-nearest neighbors
Unsupervised	inputs	characterize where the data are; explore and summarize the data	clustering, topic modeling, hidden Markov models
Reinforcement learning	inputs and delayed or partially observed outputs	maximize outputs; balance exploration and exploitation	multi-armed bandit, Markov decision process

Types of Machine Learning

Type	Observations	Gist	Examples
Supervised	inputs and outputs	learn function from input to output; maximize predictive accuracy	**logistic regression**, neural networks, K-nearest neighbors
Unsupervised	inputs	characterize where the data are; explore and summarize the data	**clustering**, topic modeling, hidden Markov models
Reinforcement learning	inputs and delayed or partially observed outputs	maximize outputs; balance exploration and exploitation	multi-armed bandit, Markov decision process

Classification

Recall Regression

- $\mathbb{E}[Y|\mathbf{X}] = f(\mathbf{X}, \beta)$
- What if the output is a discrete value (e.g., succeed or failed at exam, a link was clicked or not clicked)?
- We call this *classification*.

Classification

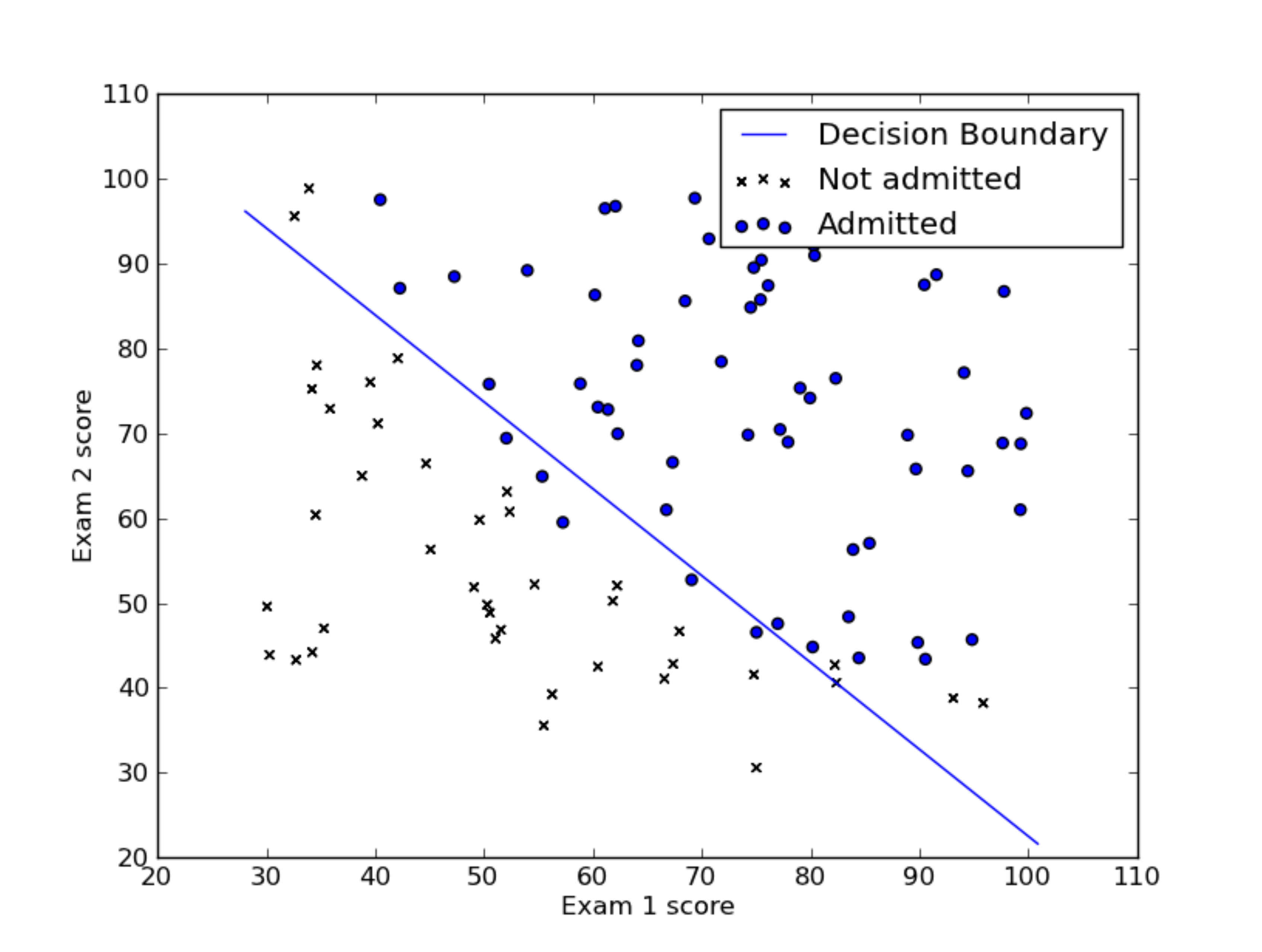
- Independent variables $\mathbf{X}$ (continuous or discrete)
- Dependent variable Y (discrete 0/1, or discrete k = 1,..,K)
- Unknown parameters β

Classification

- Independent variables $\mathbf{X}$ (continuous or discrete)
- Dependent variable Y (discrete 0/1, or discrete k = 1,..,K)
- Unknown parameters β

- For the binary case, β describes a boundary hyperplane in some way. Then, the most likely prediction is 1 above the plane and 0 below the plane.

Classification



http://stackoverflow.com/questions/30427819/plot-decision-boundary-for-scikit-logistic-regression-with-7-features

Logistic ~~Classification~~ Regression

Linear Regression	Logistic Regression
real independent variables	real independent variables
real dependent variable	discrete dependent variable
expected dependent var. is the inner product between inputs and unknown coefficients	expected dependent var. is the inner product between inputs and unknown coefficients put through a “squashing” function
closed-form solution	requires approximation or iteration

Generalized Linear Models

- Recall from linear regression document and talk:

$$\mathbb{E}[Y|\mathbf{X}] = f(\mathbf{X}, \beta)$$

$$\text{where} \quad f(\mathbf{X}, \beta) = \mathbf{X}^\top \beta$$

Generalized Linear Models

- Recall from linear regression document and talk:

$$\mathbb{E}[Y|\mathbf{X}] = f(\mathbf{X}, \beta)$$

$$\text{where } \; f(\mathbf{X}, \beta) = \mathbf{X}^\top \beta$$

- Are we constrained to this f only?

Generalized Linear Models

- Recall from linear regression document and talk:

$$\mathbb{E}[Y|\mathbf{X}] = f(\mathbf{X}, \beta)$$

$$\text{where } \; f(\mathbf{X}, \beta) = \mathbf{X}^\top \beta$$

- Are we constrained to this f only? No.

Generalized Linear Models

- Recall from linear regression document and talk:

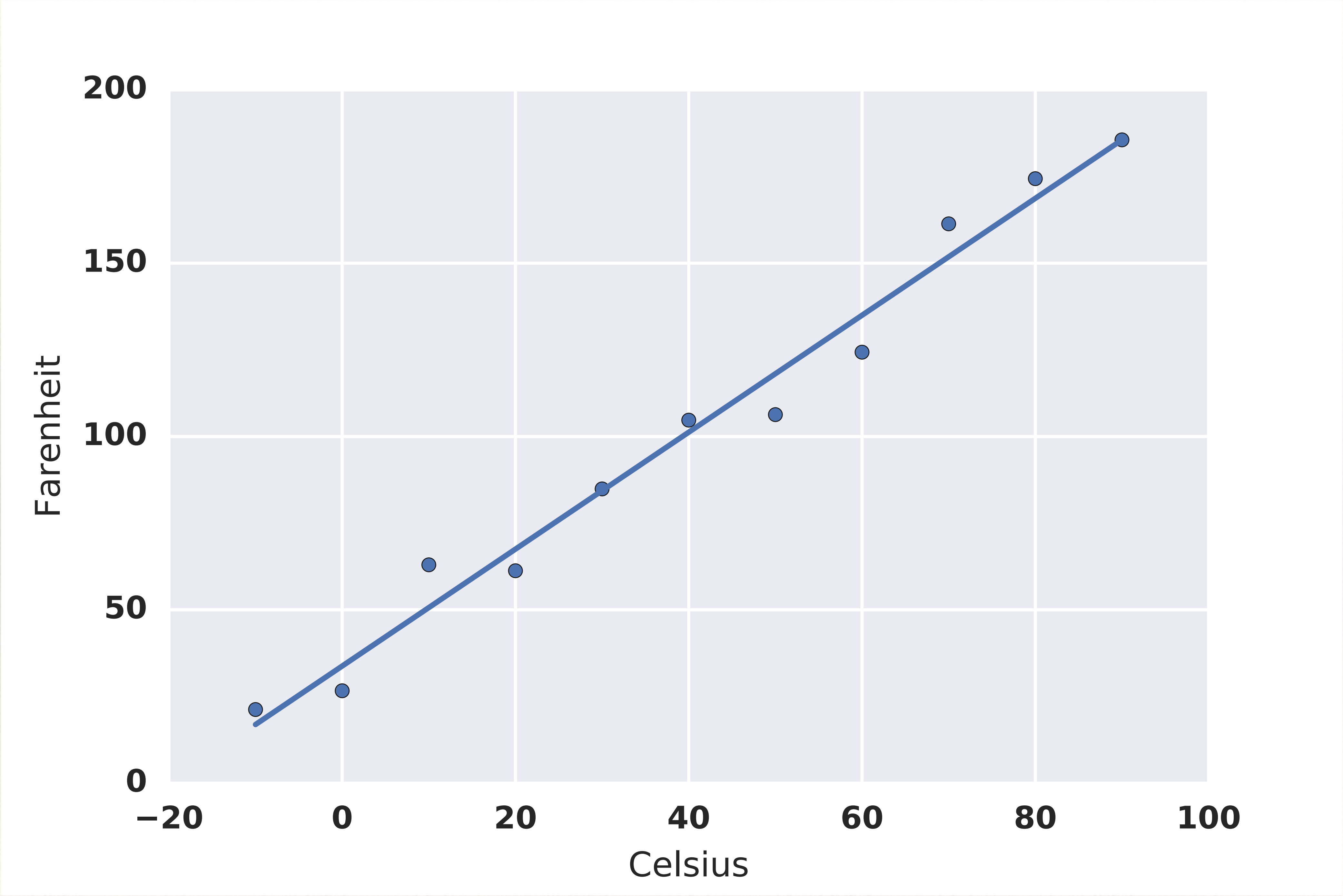
$$\mathbb{E}[Y|\mathbf{X}] = f(\mathbf{X}, \beta)$$

$$\text{where} \quad f(\mathbf{X}, \beta) = \mathbf{X}^\top \beta$$

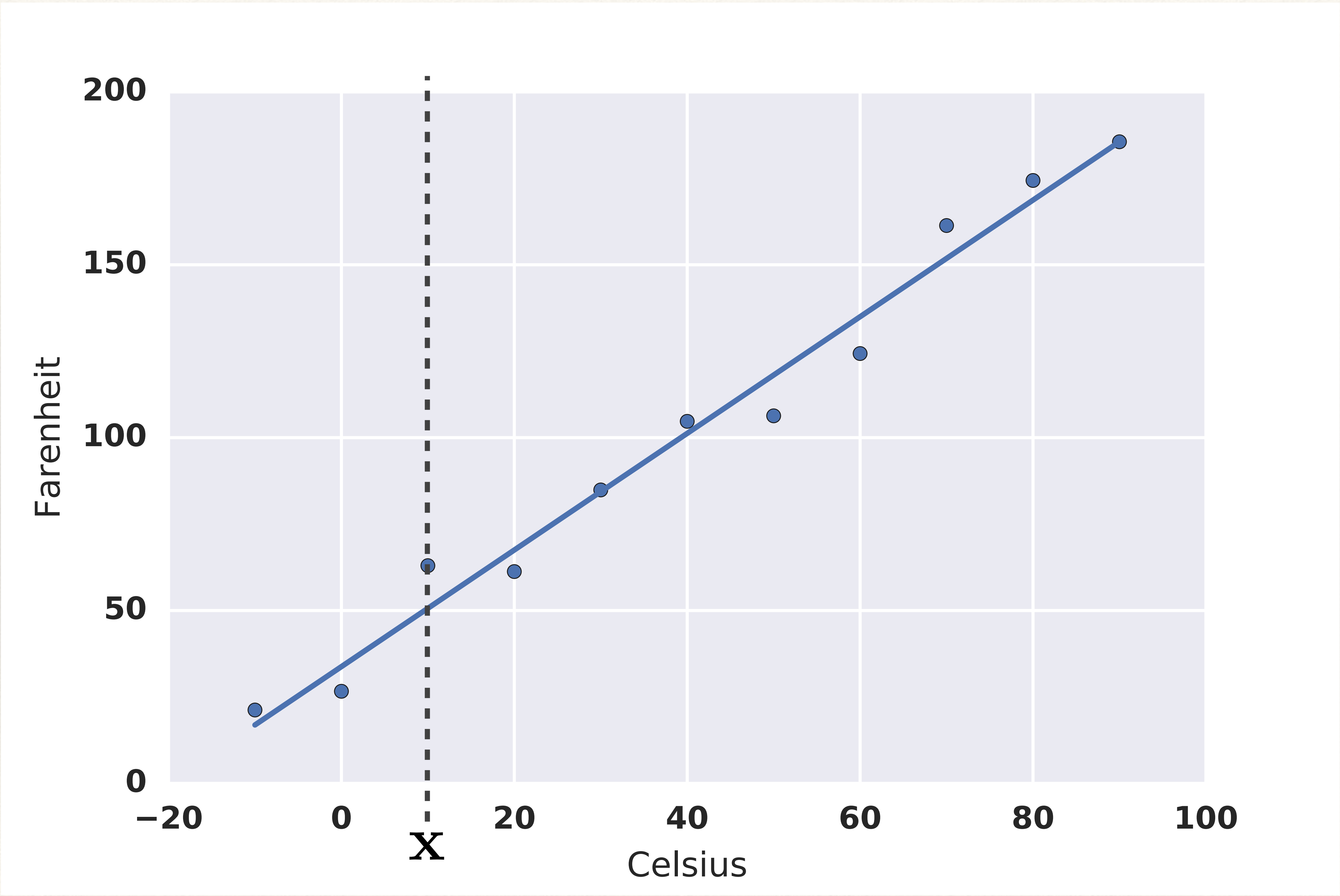
- Are we constrained to this f only? No.

- Generalized linear models (GLM) allow us to model many different types of data (continuous, categorical, counts) using the same framework: "all" we have to do is vary f and the distribution of the dependent variable.

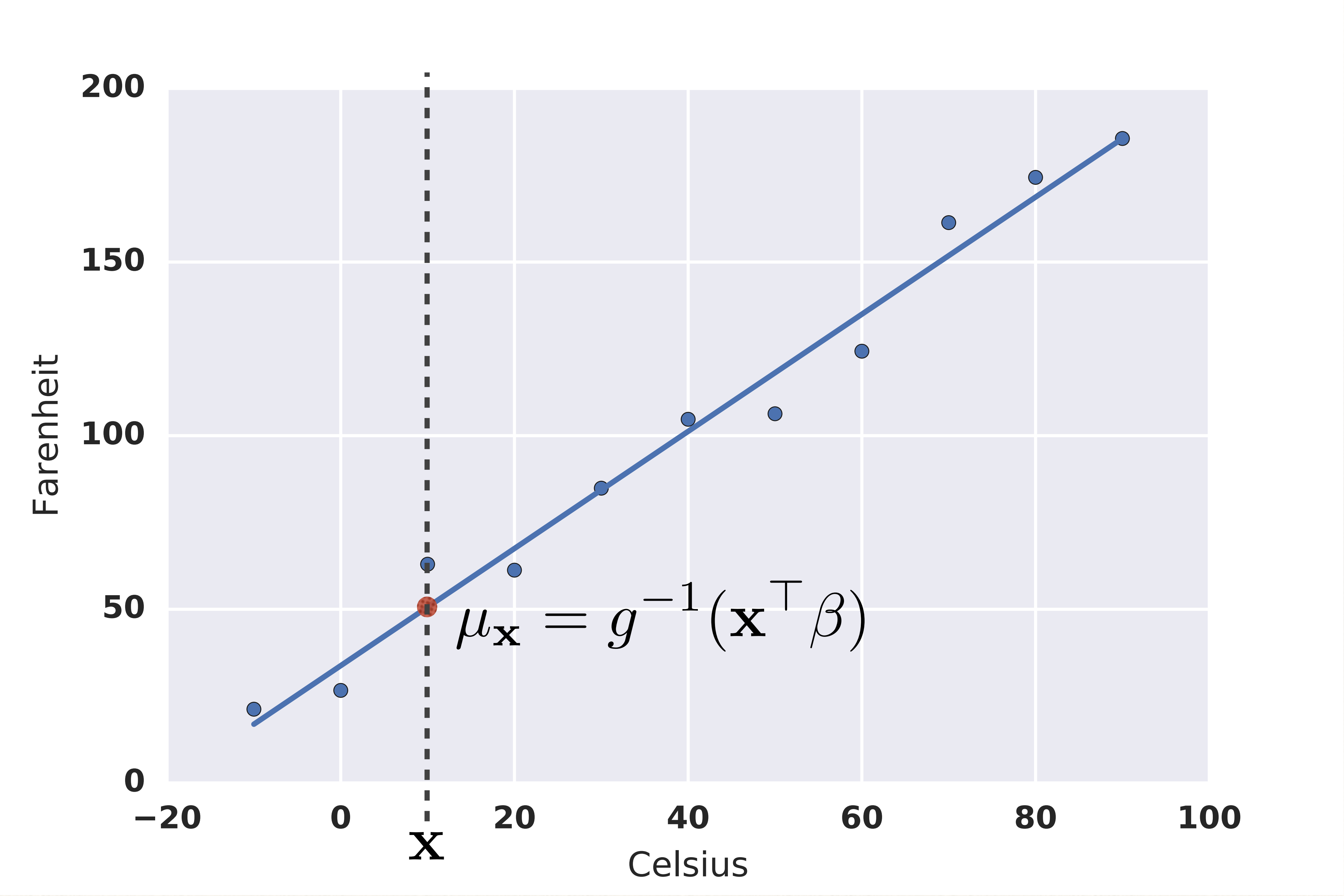
Probability Distribution of the Dependent Variable



Probability Distribution of the Dependent Variable

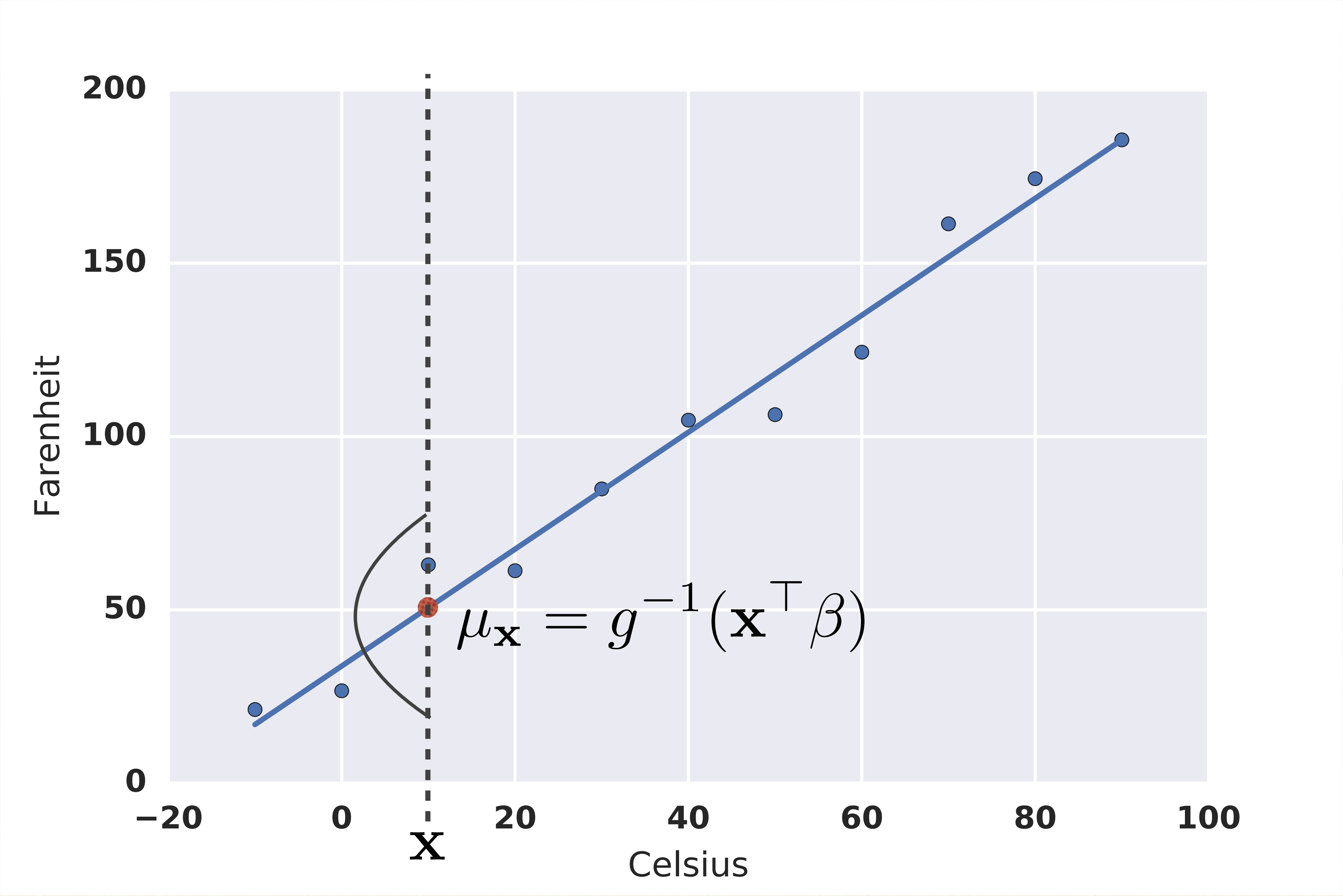


Probability Distribution of the Dependent Variable

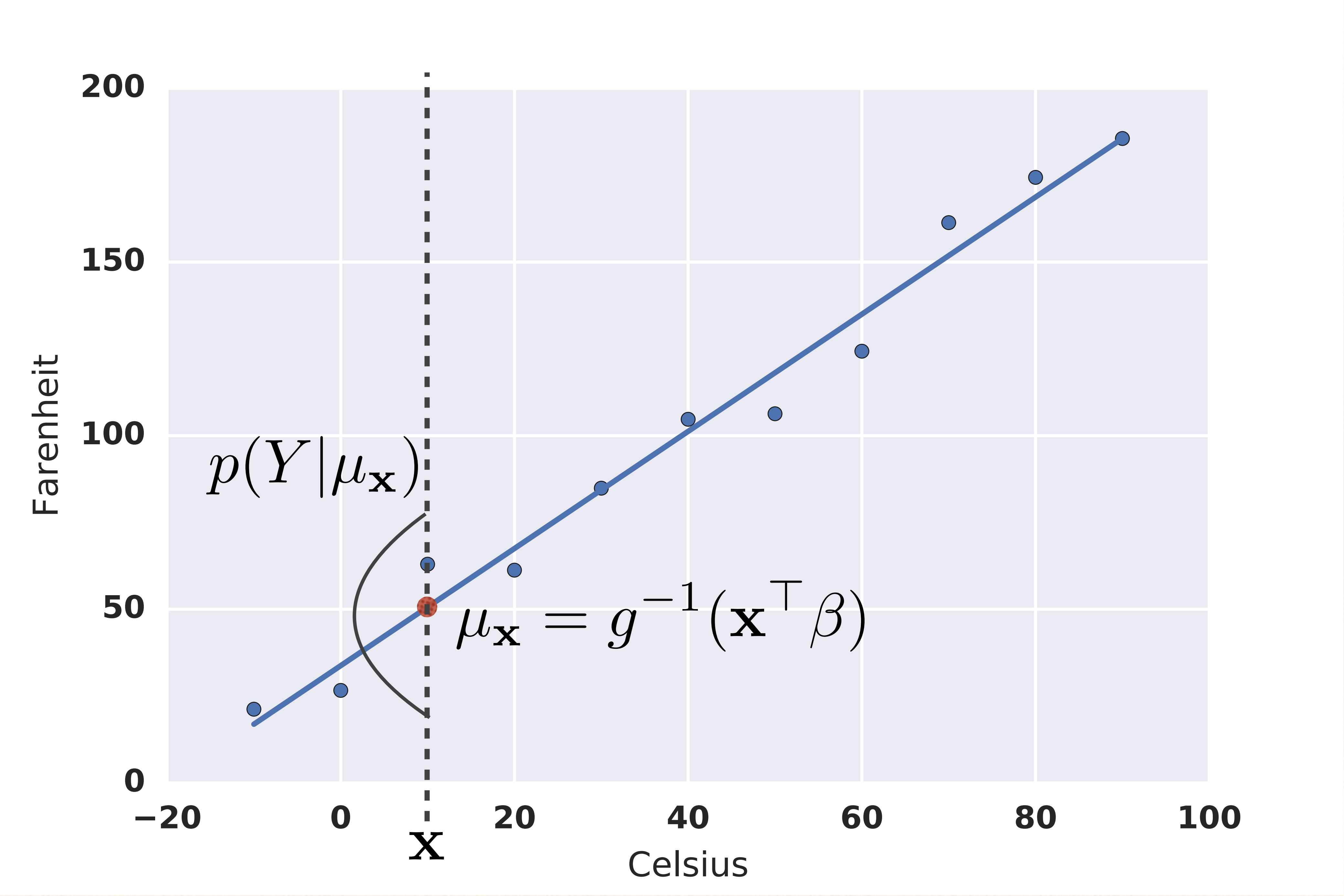


COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Probability Distribution of the Dependent Variable



Probability Distribution of the Dependent Variable



Generalized Linear Models

- Putting it all together, the GLM is a family of models specified by:
  - link function g;
  - probability distribution of dependent var. $p(Y|\mu_{\mathbf{x}})$;
  - linear predictor $\eta = \mathbf{X}^\top \beta$. (Puts the "L" in GLM).

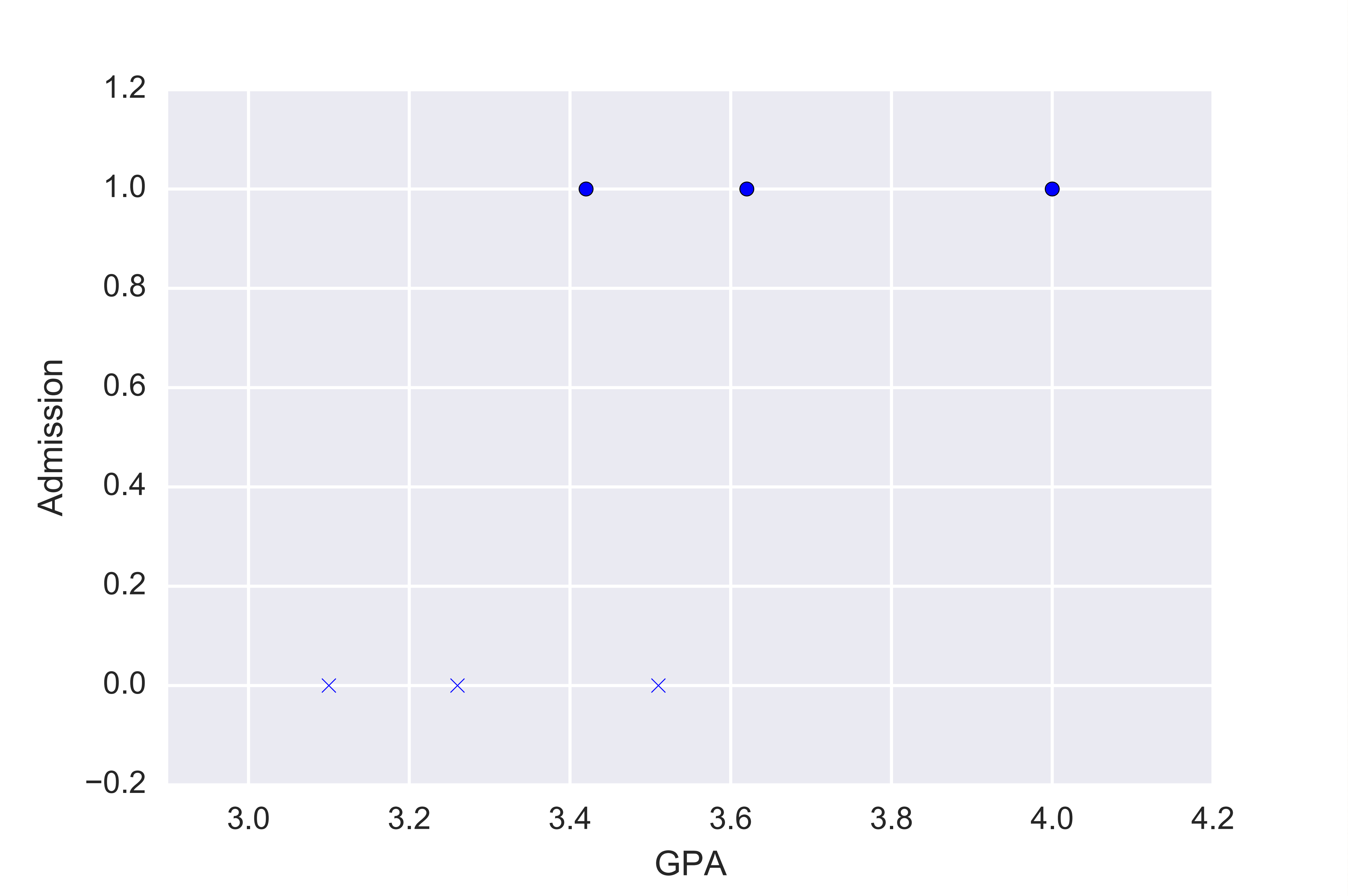
Generalized Linear Models

- Putting it all together, the GLM is a family of models specified by:
  - link function g;
  - probability distribution of dependent var. $p(Y|\mu_{\mathbf{x}})$;
  - linear predictor $\eta = \mathbf{X}^\top \beta$. (Puts the "L" in GLM).
- Logistic regression is a type of GLM…

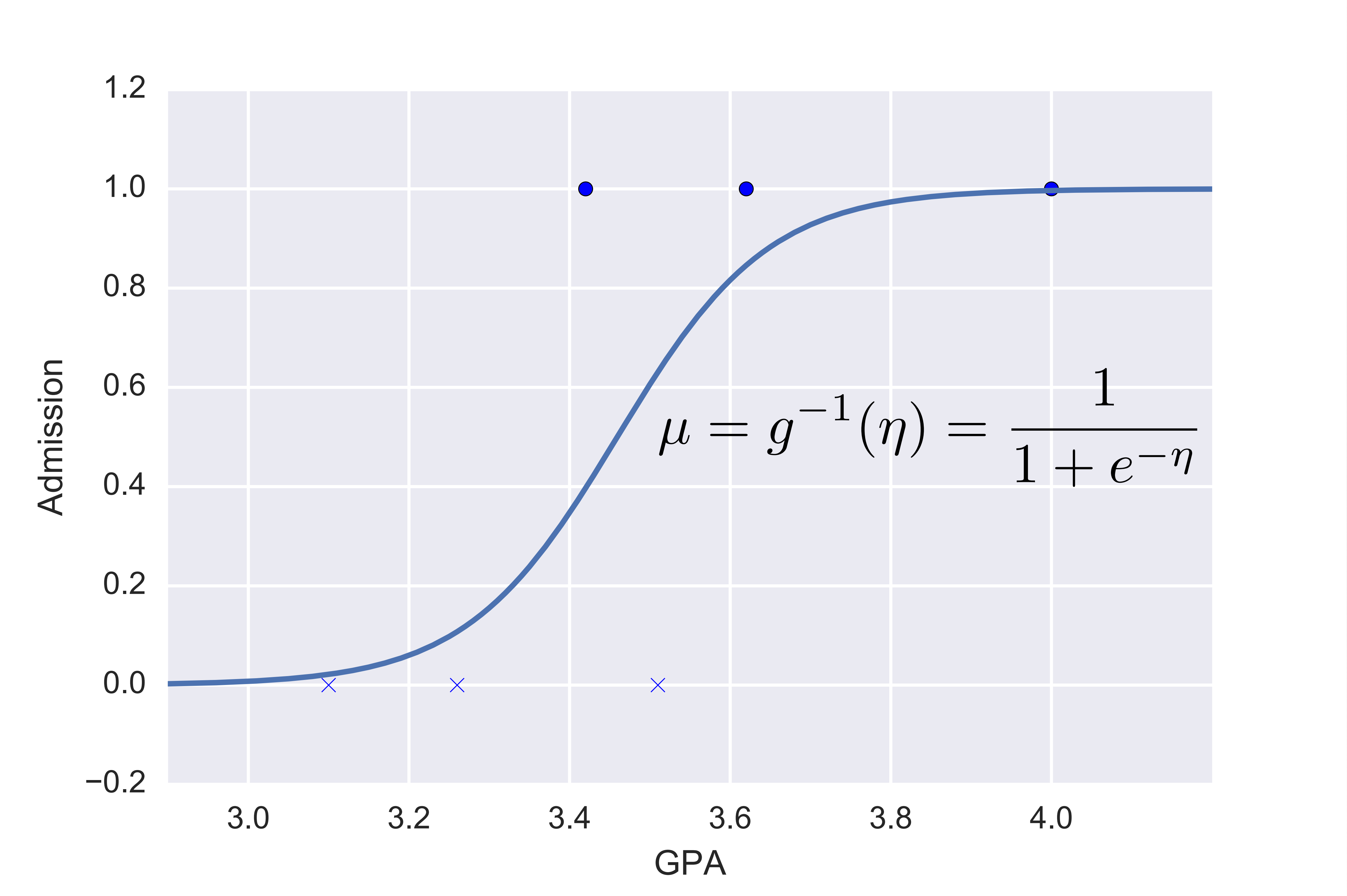
Logistic Regression

- Logistic regression is a type of GLM with:
  - link function $\eta = g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
  - implies: inverse link function $\mu = g^{-1}(\eta) = \frac{1}{1+e^{-\eta}}$
  - Bernoulli $p(Y|\mu_{\mathbf{x}}) = \mathcal{B}(Y|\mu_{\mathbf{x}})$

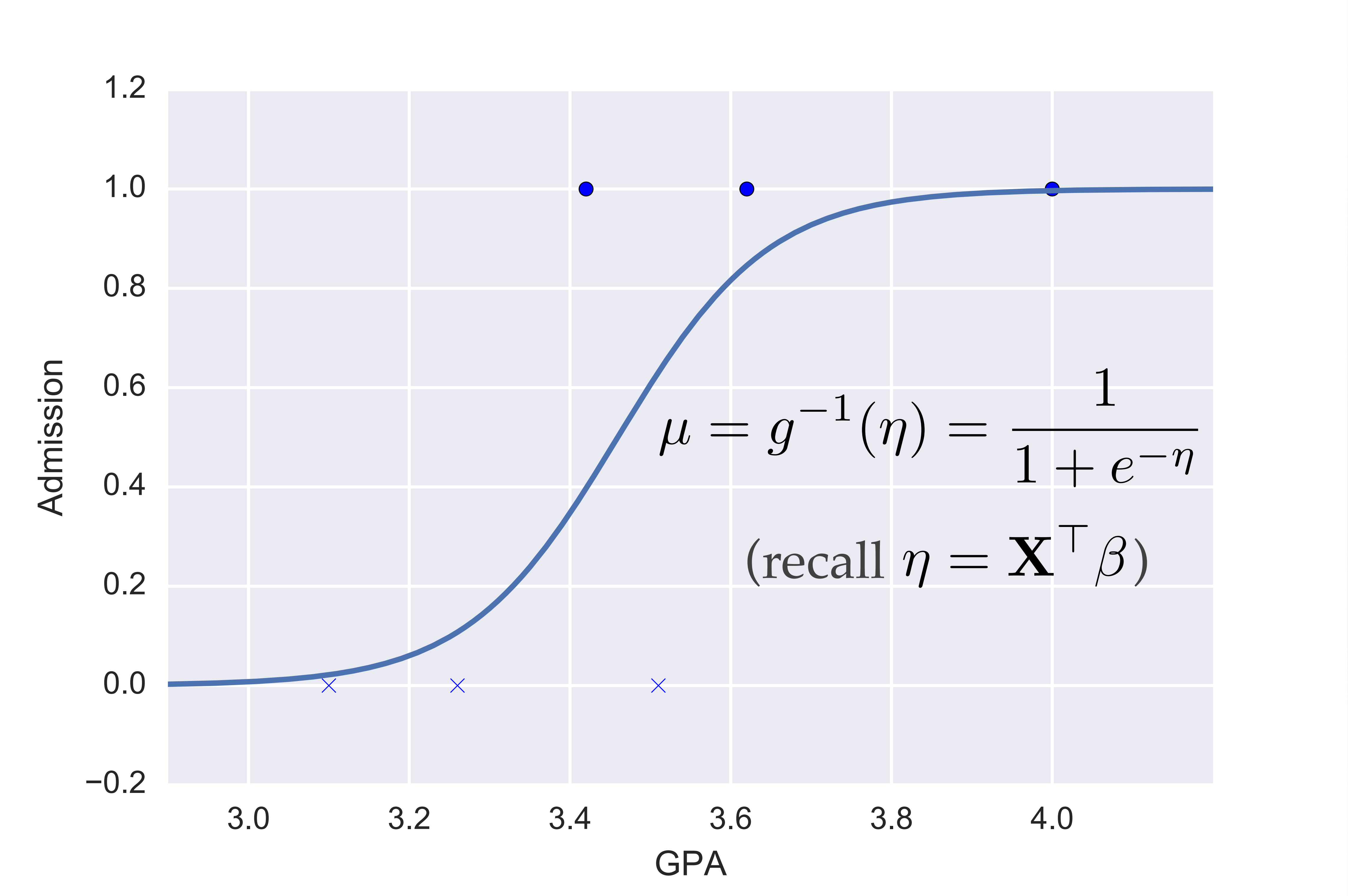
Logistic Regression



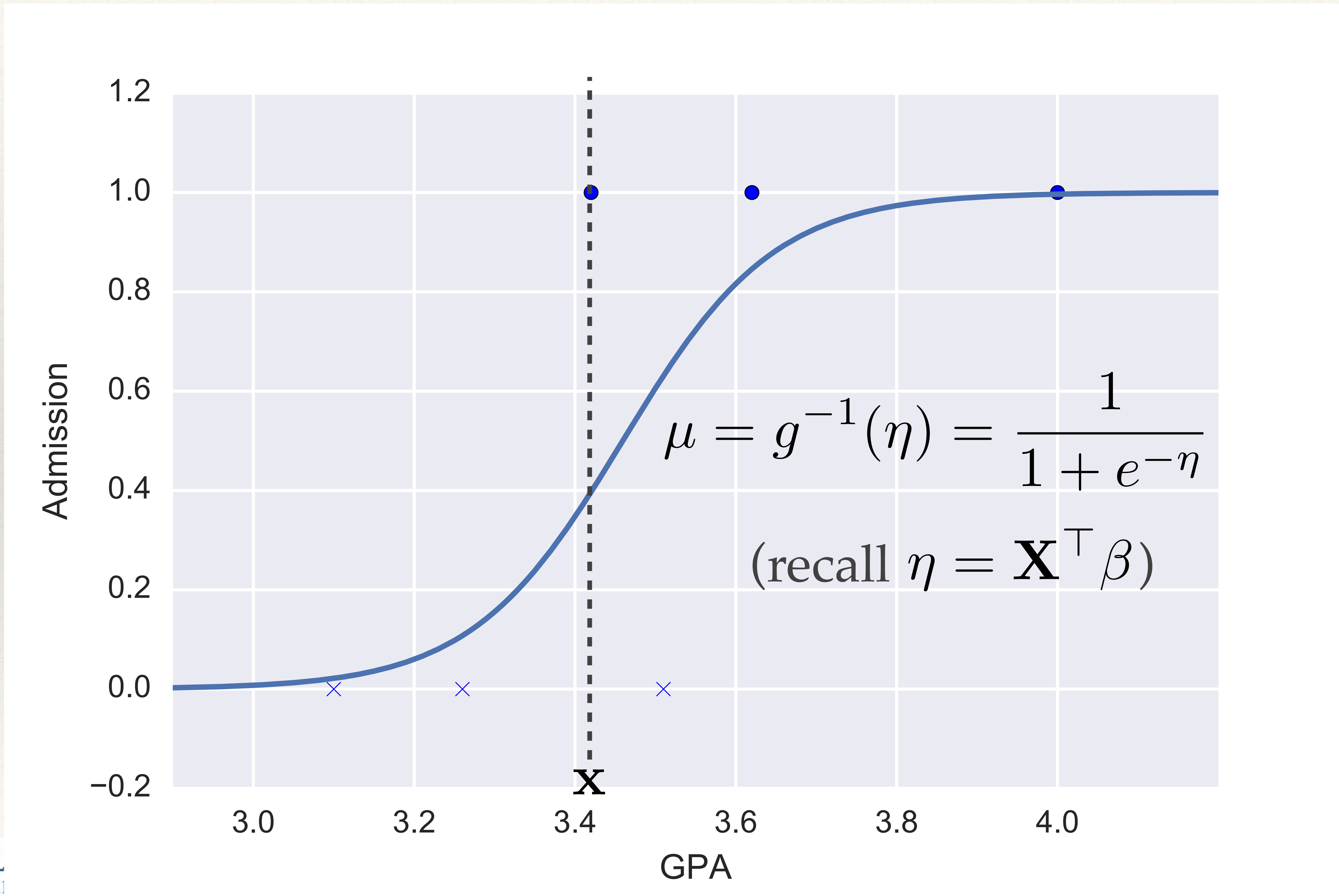
Logistic Regression



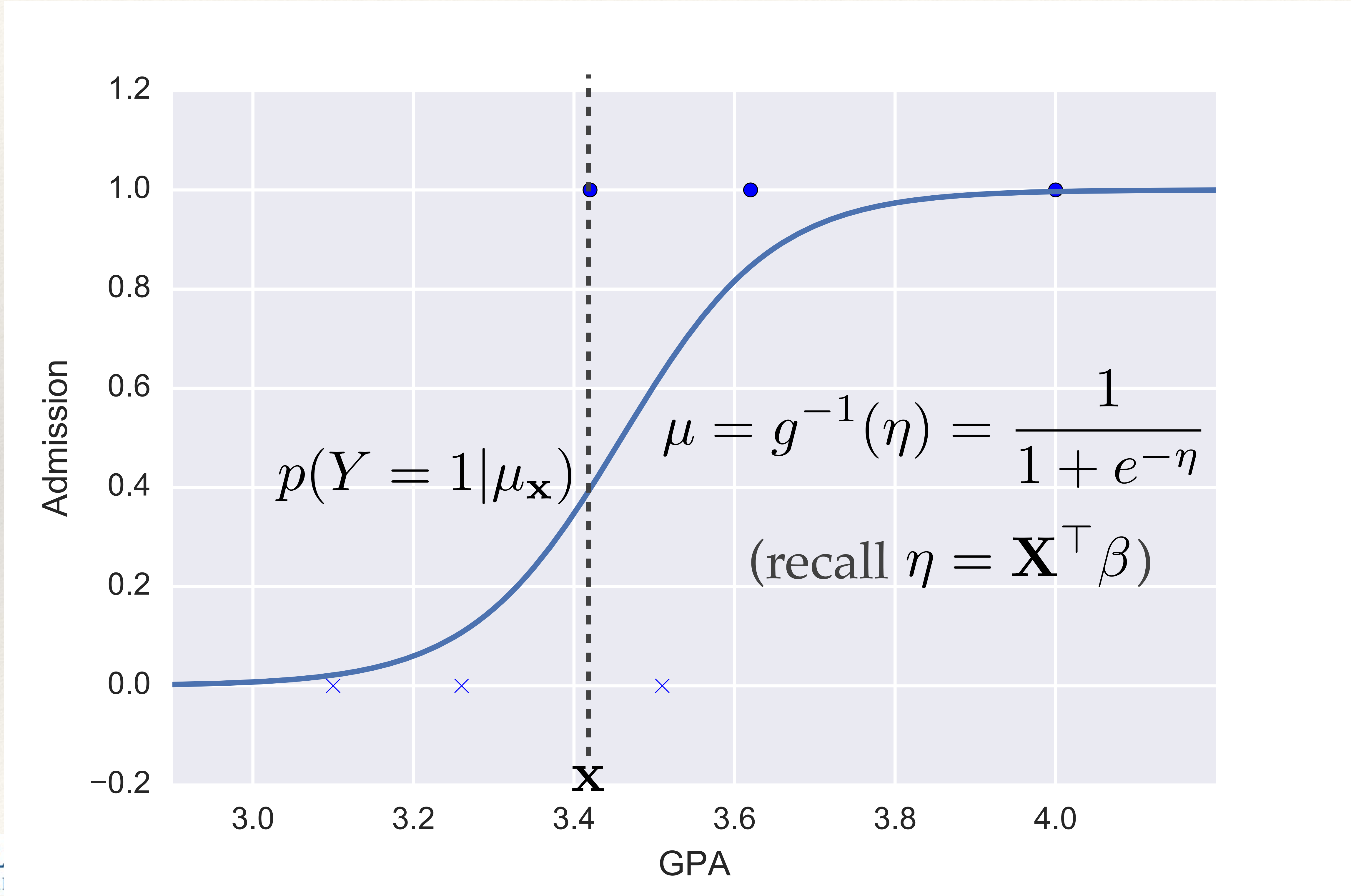
Logistic Regression



Logistic Regression

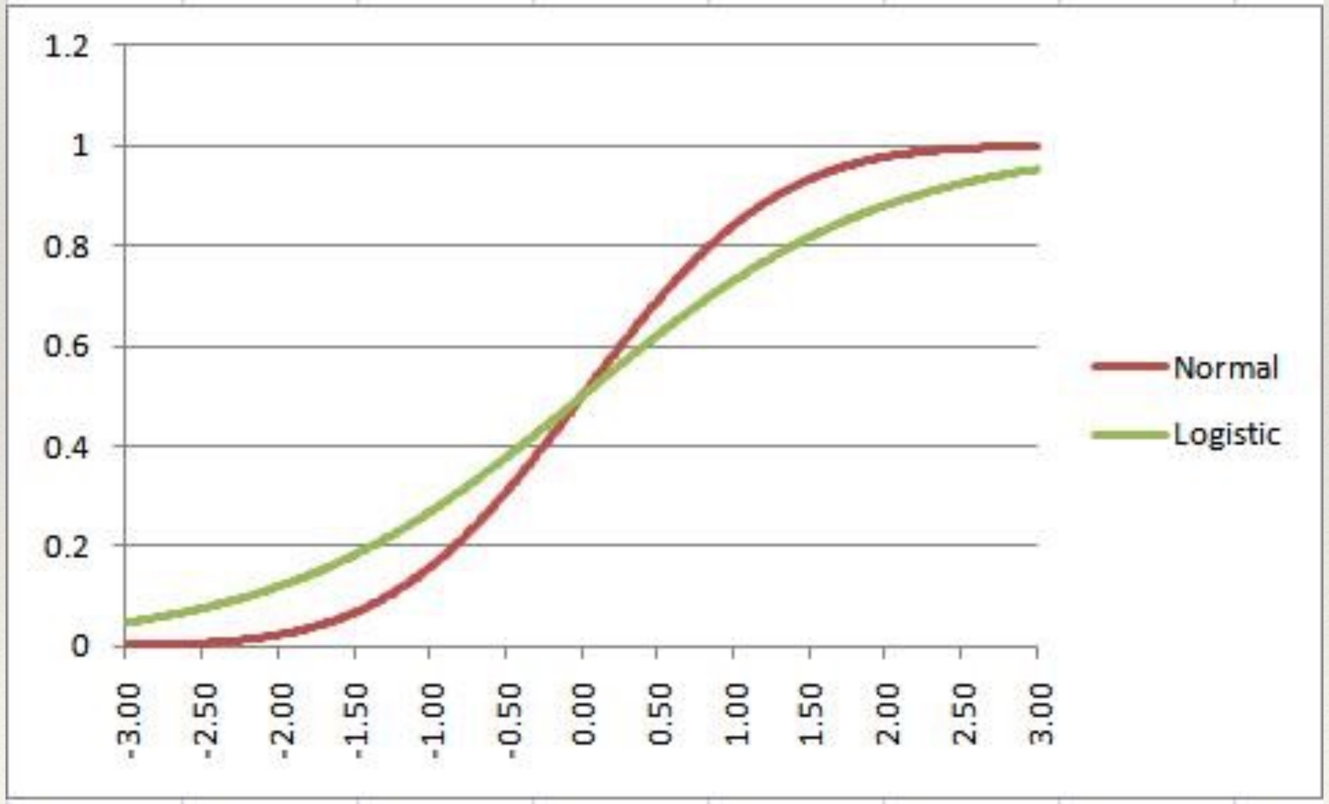


Logistic Regression



"Squashing" Functions

- Here are some other "squashing" functions you are likely to encounter in a classification setting:
  - Probit vs. logistic:

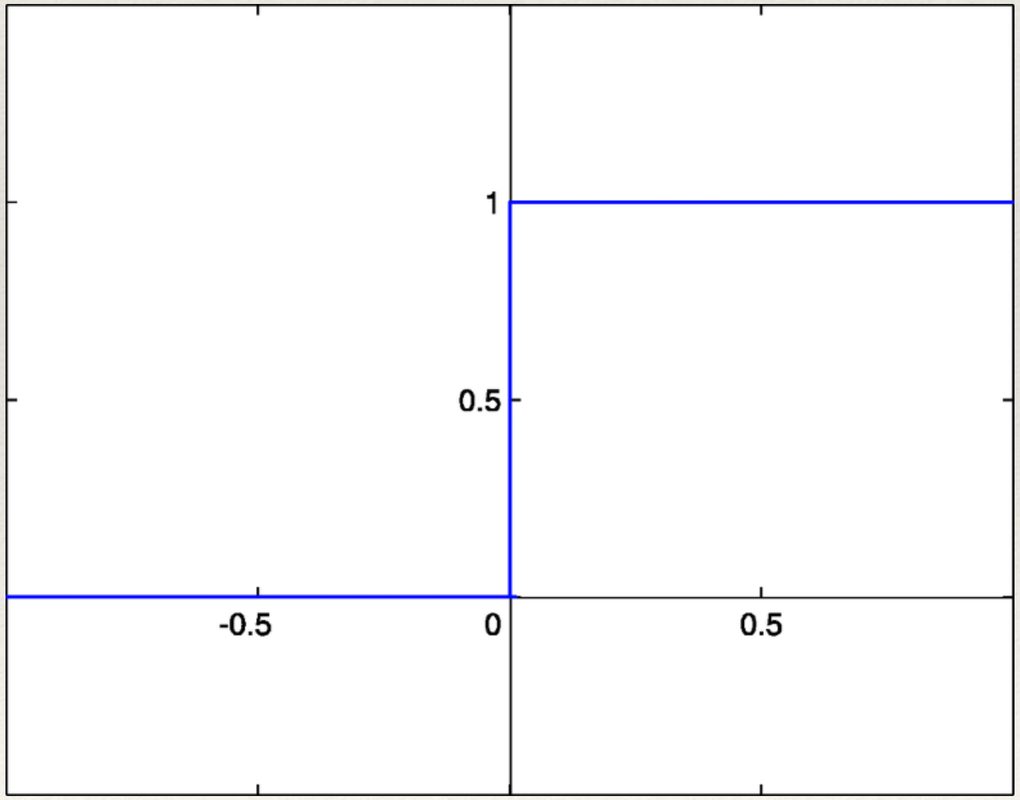


COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

from: http://dss.princeton.edu/online_help/stats_packages/stata/probit.html

"Squashing" Functions

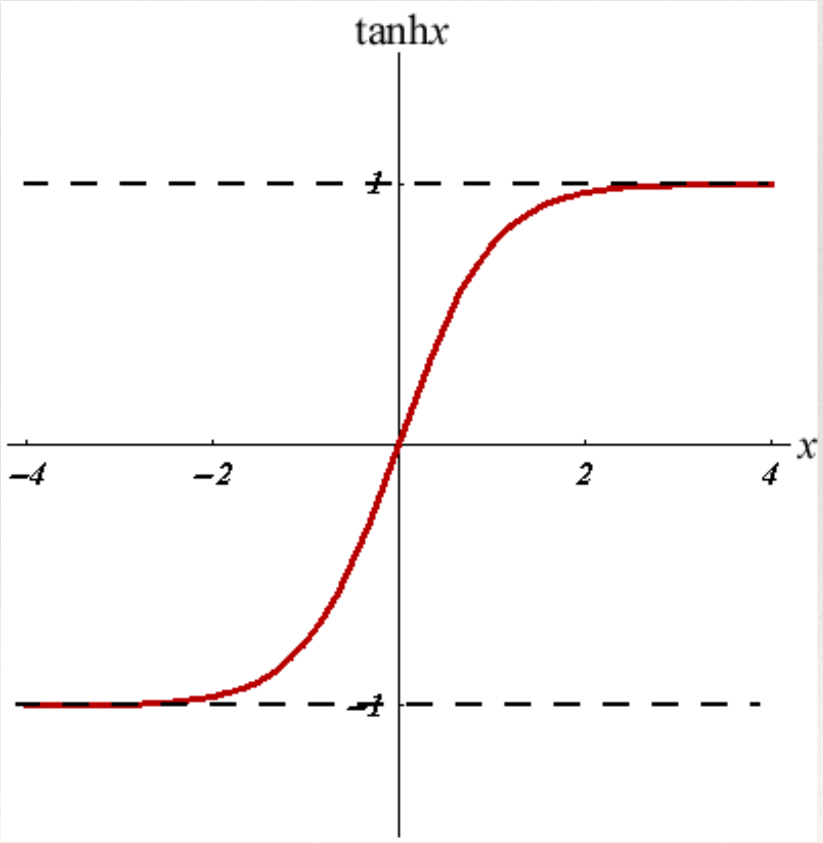
- Here are some other "squashing" functions you are likely to encounter in a classification setting:
  - Step function:



from: https://en.wikibooks.org/wiki/Artificial_Neural_Networks/Activation_Functions

“Squashing” Functions

- Here are some other “squashing” functions you are likely to encounter in a classification setting:
  - Hyperbolic tangent:



from: http://www.efunda.com/math/hyperbolic/display.cfm?name=tanh

Training Logistic Regression

- Not as straightforward as for linear regression because there is no closed-form solution for the coefficients.
- Usually resort to a maximum likelihood or Bayesian Markov chain Monte Carlo simulation.
- Maximum likelihood optimizes an objective function using gradient descent (covered this afternoon).

Clustering

Clustering

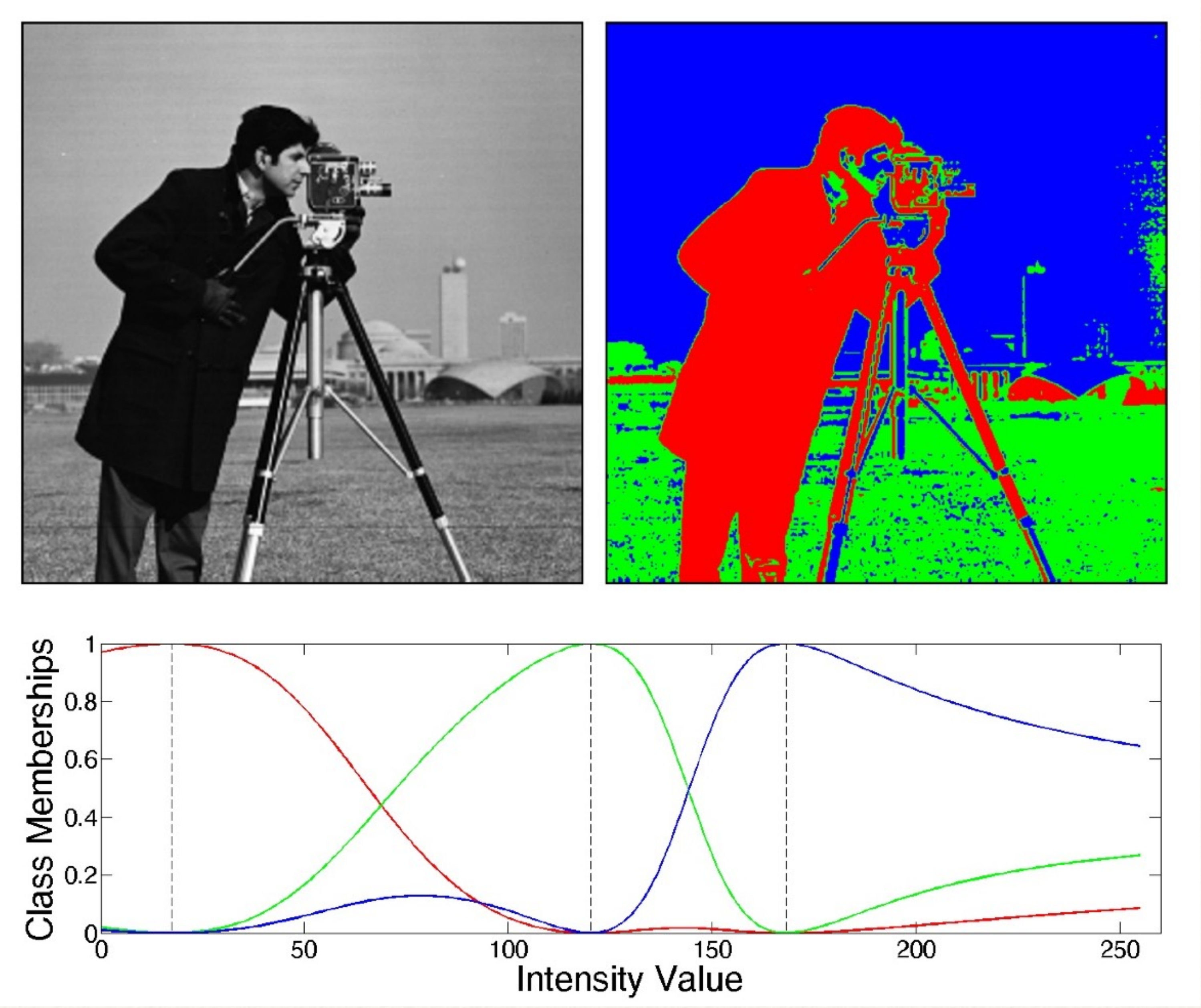
Classification	Clustering
real independent variables	real independent variables
discrete dependent variable	no dependent variable
supervised learning	unsupervised learning
error based on predicted and observed dependent variable	error based on squared distance to nearest cluster center

Why Cluster?

- Find groupings in the data in an unsupervised way;
- characterize data that may be hard to plot or read (i.e., > 3 dimensions or large data set);
- useful for initializations for some more complicated model (e.g., factorial mixtures, HMMs).

Examples of Clustering

Clustering for image segmentation



https://www.mathworks.com/matlabcentral/fileexchange/41967-fast-segmentation-of-n-dimensional-grayscale-images

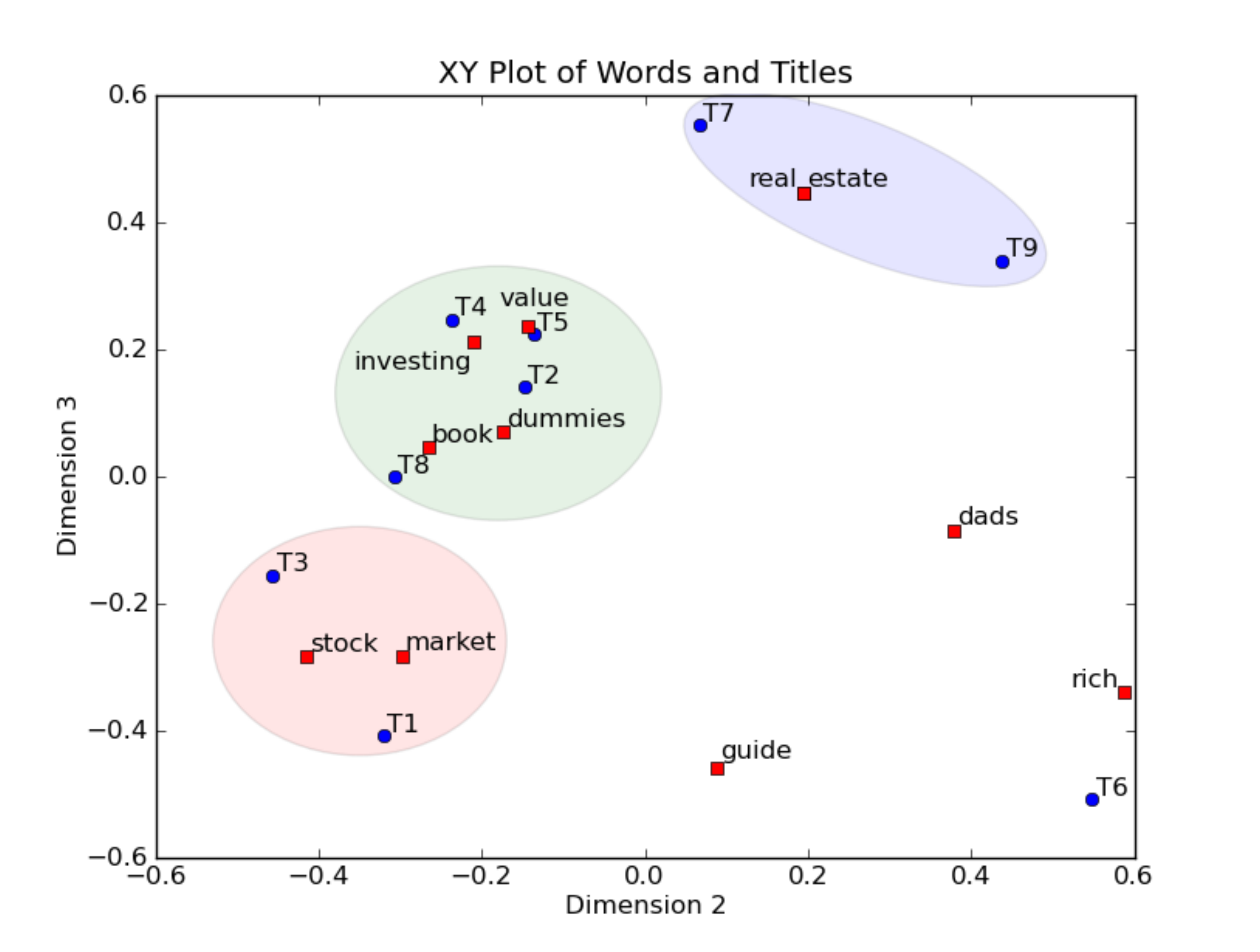
Examples of Clustering

Clustering for image compression

Bishop, Pattern Recognition & Machine Learning, 2006

Examples of Clustering

Clustering documents of text

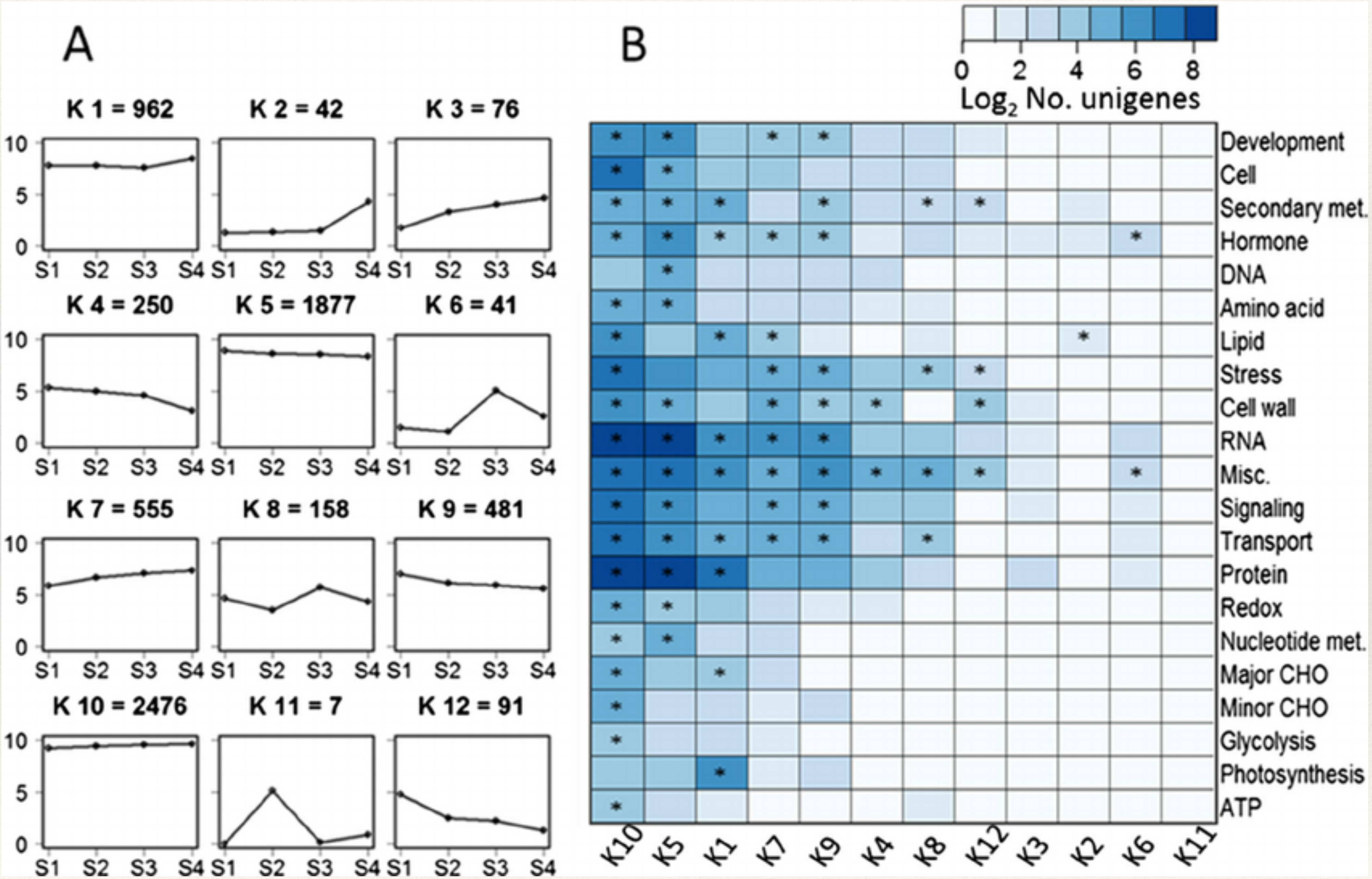


https://technowiki.wordpress.com/2011/08/27/latent-semantic-analysis-lsa-tutorial/

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

Examples of Clustering

Clustering genes



https://www.researchgate.net/publication/267341822_Transcriptome_Analysis_of_European_Pears_Pyrus_communis_cv_Bartlett_Reveals_Significant_Gene_Families_and_Co-expressed_Genes_Associated_with_Development_of_Ripening_Capacity_by_Temperature_Conditionin

K-Means Clustering

- K-means is the most popular clustering algorithm used all across in academia and industry.
- It is an iterative procedure that alternates between updating cluster centres for $k = 1,..,K$ and assignments of data point $n = 1,..,N$ to cluster $k = 1,..,K$.
- The procedure will converge to a local minimum that minimizes the sum of squared distance to nearest cluster centre for each data point.

K-Means Clustering

- Pseudocode for iterative procedure:

1. randomly initialize μ_k, for $k = 1, \ldots, K$
2. while μ not converged:
   A. assign data point n to nearest cluster:
   $r_n \leftarrow \arg\min_k ||x_n - \mu_k||^2$, for $n = 1, \ldots, N$
   B. count number of data points assigned to cluster k,
   $N_k \leftarrow \sum_{n=1}^{N} r_{nk}$, for $k = 1, \ldots, K$
   C. update cluster centers: $\mu_k \leftarrow \frac{1}{N_k} \sum_{n=1}^{N} x_n r_{nk}$, for $k = 1, \ldots, K$

Evaluating Cluster Results

$$\mathcal{L}(R) = -\sum_{n=1}^{N}\sum_{k=1}^{K} R_{nk}||X_n - \mu_k||^2$$

- "Find the sum of squared distances to the nearest cluster centre for each data point".
- Cross-validation.
- Visualize!

Choosing K

- Similar to the overfitting dilemma of regression:
  - a too-large K will overfit the data ("one cluster per data point");
  - a too-small K will not fit the data enough.
- Methods:
  - use domain knowledge;
  - error analysis with different K;
  - add penalty for new clusters: BIC, AIC, Dirichlet process;
  - model-based methods, using a Bayesian prior.

Limitations of K-Means

- Sensitive to outliers.
- Clusters are always spheres.
- Converges to a **local minimum** of the objective function
    - —> sometimes converges to a bad local minimum, need to try different random initializations.

Today's Lab Session

- Logistic regression:
  - find the unknown coefficients;
  - apply to mushroom data set.
- Clustering:
  - implement K-means from pseudocode;
  - explore limitations;
  - apply to New York City collisions data.

Conclusion

- High level view of machine learning: supervised learning, unsupervised learning, reinforcement learning.
- Introduced logistic regression: a popular classification method.
- Introduced K-means clustering: a popular clustering method that groups data in an unsupervised manner.