STEVENS INSTITUTE of TECHNOLOGY
School of Business

BIA-652

# Multivariate Data Analytics

## Classification:
## Discriminant Analysis and Logistic Regression

Prof. Feng Mai
School of Business

For academic use only.

# More on generative vs discriminative classifiers

- **Easy to fit?** As we have seen, it is usually very easy to fit generative classifiers. For example, in Sections 3.5.1.1 and 4.2.4, we show that we can fit a naive Bayes model and an LDA model by simple counting and averaging. By contrast, logistic regression requires solving a convex optimization problem (see Section 8.3.4 for the details), which is much slower.
- **Fit classes separately?** In a generative classifier, we estimate the parameters of each class conditional density independently, so we do not have to retrain the model when we add more classes. In contrast, in discriminative models, all the parameters interact, so the whole model must be retrained if we add a new class. (This is also the case if we train a generative model to maximize a discriminative objective Salojarvi et al. (2005).)
- **Handle missing features easily?** Sometimes some of the inputs (components of $\mathbf{x}$) are not observed. In a generative classifier, there is a simple method for dealing with this, as we discuss in Section 8.6.2. However, in a discriminative classifier, there is no principled solution to this problem, since the model assumes that $\mathbf{x}$ is always available to be conditioned on (although see (Marlin 2008) for some heuristic approaches).
- **Can handle unlabeled training data?** There is much interest in **semi-supervised learning**, which uses unlabeled data to help solve a supervised task. This is fairly easy to do using generative models (see e.g., (Lasserre et al. 2006; Liang et al. 2007)), but is much harder to do with discriminative models.

Murphy (2012) Probabilistic Machine Learning: An Introduction

# More on generative vs discriminative classifiers

- **Symmetric in inputs and outputs?** We can run a generative model "backwards", and infer probable inputs given the output by computing $p(\mathbf{x}|y)$. This is not possible with a discriminative model. The reason is that a generative model defines a joint distribution on $\mathbf{x}$ and $y$, and hence treats both inputs and outputs symmetrically.
- **Can handle feature preprocessing?** A big advantage of discriminative methods is that they allow us to preprocess the input in arbitrary ways, e.g., we can replace $\mathbf{x}$ with $\boldsymbol{\phi}(\mathbf{x})$, which could be some basis function expansion, as illustrated in Figure 8.9. It is often hard to define a generative model on such pre-processed data, since the new features are correlated in complex ways.
- **Well-calibrated probabilities?** Some generative models, such as naive Bayes, make strong independence assumptions which are often not valid. This can result in very extreme posterior class probabilities (very near 0 or 1). Discriminative models, such as logistic regression, are usually better calibrated in terms of their probability estimates.

Murphy (2012) Probabilistic Machine Learning: An Introduction

# Confusion Matrix and Decision Threshold

# Evaluation of Classification Model's Performance

Example:

We train a logistic regression model on a train set to predict loan defaults for a bank.

We apply the trained model on a test set (10000 obs). The logistic regression model outputs ln(Odds), which can be converted to predicted P(Default | X).

We need to set a Decision Threshold or Cutoff Probability: If P(Default | X) > 0.5 → Predict Default = Yes

We compare the predicted classification with the true Y label in the test set using a Confusion Table

Correct

Incorrect

| | | *True Default Status* | | |
|---|---|---|---|---|
| | | No | Yes | Total |
| *Predicted Default Status* | No | 9644 | 252 | 9896 |
| | Yes | 23 | 81 | 104 |
| | Total | 9667 | 333 | 10000 |

Misclassification Rate = (252+23)/10000 = 2.75%

Which is more costly?

False negative (FN) rate = FN/P = 252/333 = 75.5%

False positive (FP) rate = FP/N = 23/9667 = 0.237%

Source: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

# Adjusting Decision Threshold (from 0.5 to 0.2)

- If P(Default | X) > 0.2 → Predict Default.
- More applicants will be predicted as Default = Yes

| | | *True Default Status* | | |
|---|---|---|---|---|
| | | No | Yes | Total |
| *Predicted Default Status* | No | 9432 | 138 | 9570 |
| | Yes | 235 | 195 | 430 |
| | Total | 9667 | 333 | 10000 |

Misclassification Rate = (235+138)/10000 = 3.73% ↑

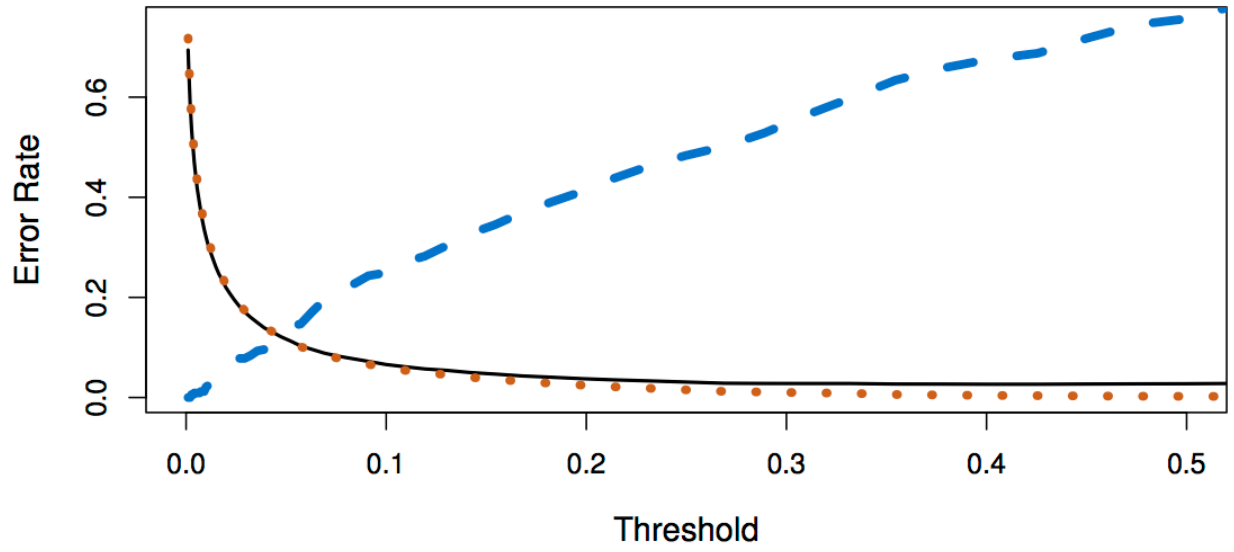
False negative (FN) rate = FN/P = 138/333 = 41.4% ↓ From 75.5%

False positive (FP) rate = FP/N = 235/9667 = 2.4% ↑

Is it worth it? Overall Cost = $(FN) + $(FP)

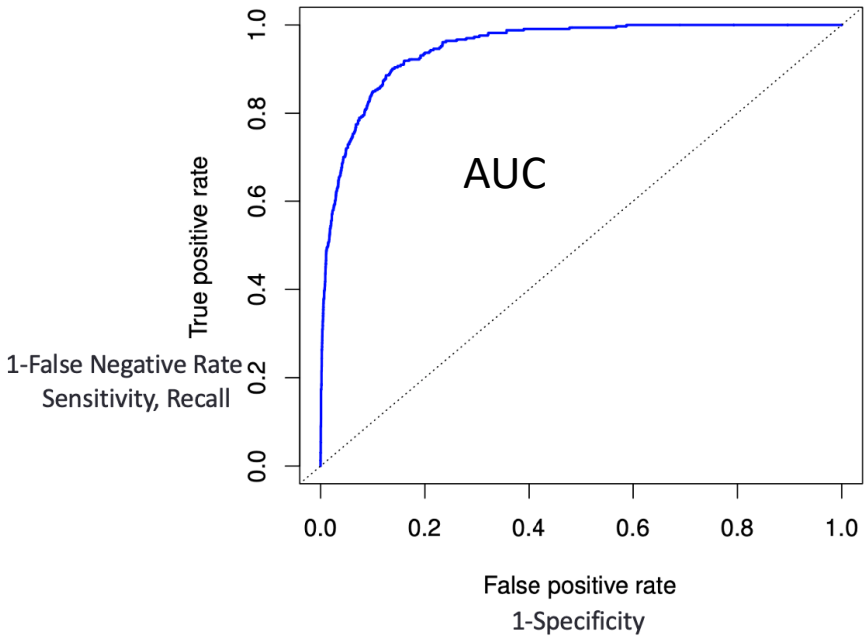
# Threshold Values (Cut-off Probability) vs. Error Rates

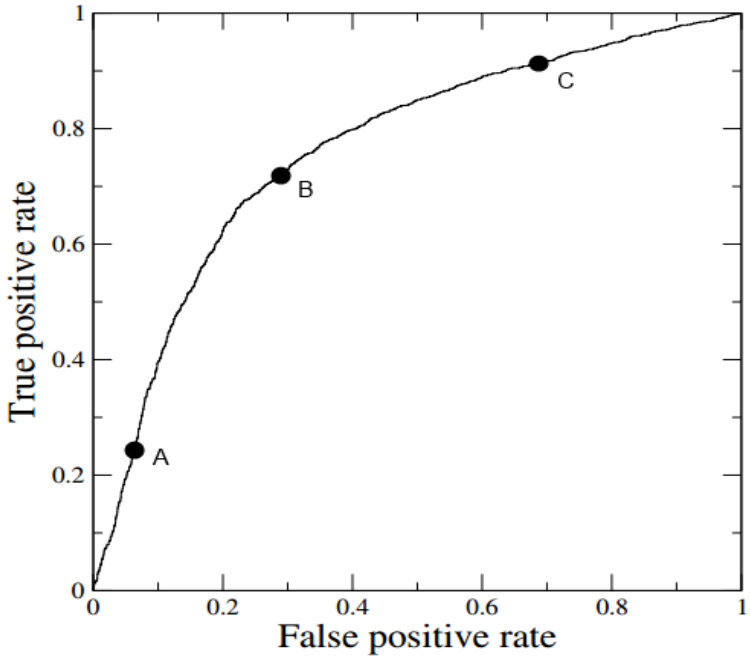
- Black solid: overall error rate
- Blue dashed: Fraction of defaulters missed (FN)
- Orange dotted: non defaulters incorrectly classified as default (FP)



# ROC (Receiver Operating Characteristics) Curve

- The ROC curve displays the trade-off between FP and FN when decision threshold changes.
- Given the same classification model, a unique decision threshold between [0, 1] corresponds to a point on the ROC curve.
- We can use the AUC or area under the curve to summarize the overall performance. Higher AUC is better.
- AUC could be a better measure than misclassification rate or accuracy when the cost is asymmetric






Suppose cut-off probability values of 0.2, 0.4, and 0.9 are marked on the curve.

Which one of A, B, C corresponds to cut-off probability 0.9?

# Multivariate Data Analytics

## Regression and Causality

Prof. Feng Mai
School of Business

For academic use only.

# Correlation != Causation

**A Simple Trick to Make Your Home Worth $6,000 More**

You don't need a touch-screen refrigerator or a Wi-Fi-enabled thermostat to pump up the sale price of your home. In fact, all you need is a black door.

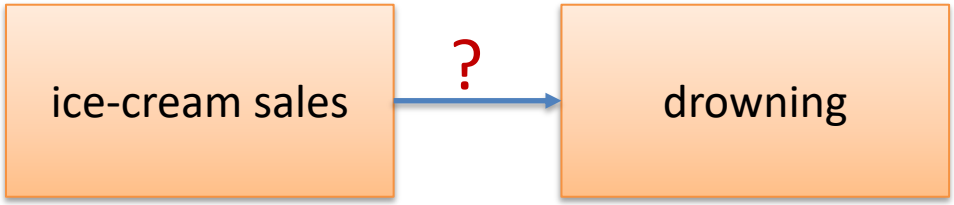
That's according to the online real estate database Zillow, which examined some 135,000 photos from listings across the U.S. since 2010. The company found that on average, houses with black or charcoal-gray front doors sold for as much as $6,271 more than expected.

Sales Price = $b_0 + b_1$*Sqft $+b_2$*Rooms + $b_3$*lotsize + $b_4$*blackdoor + ...


Source: Money Magazine, October 2018

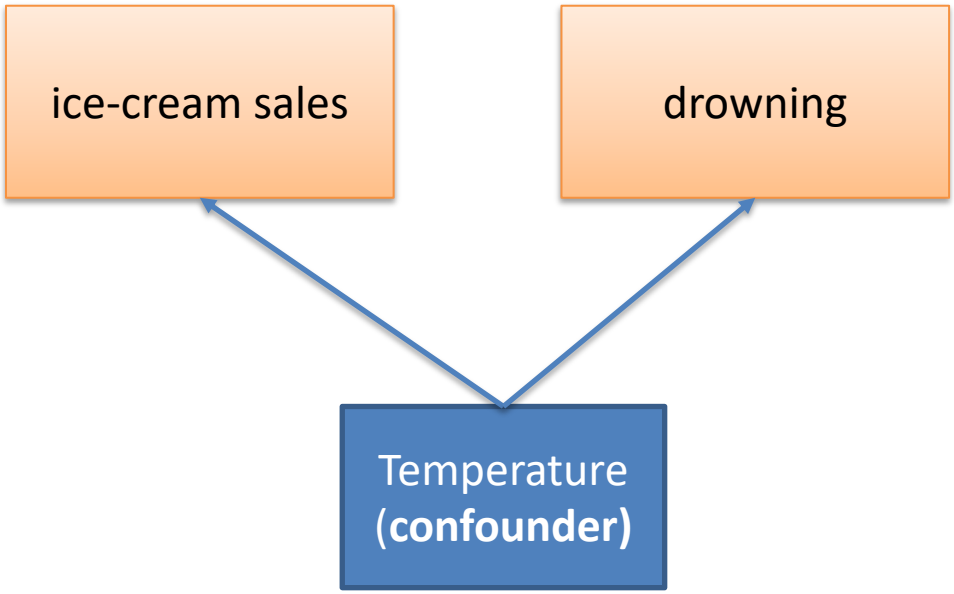
# Confounder

- Days with more ice-cream sales have more drowning accidents.

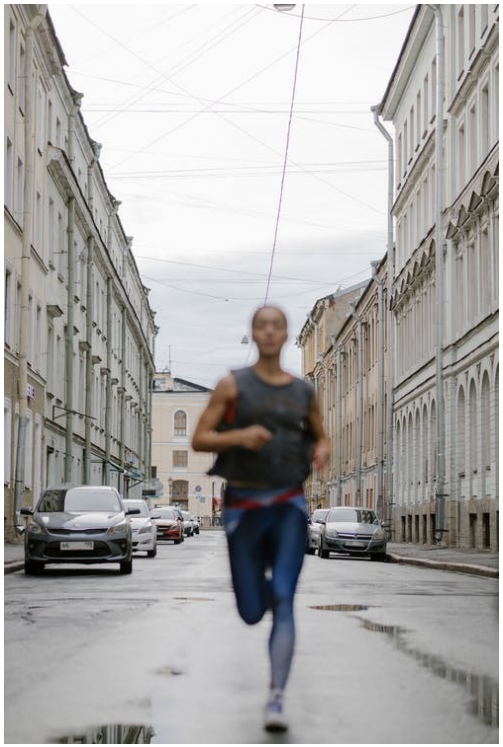


# Confounder

- Days with more ice-cream sales have more drowning accidents.



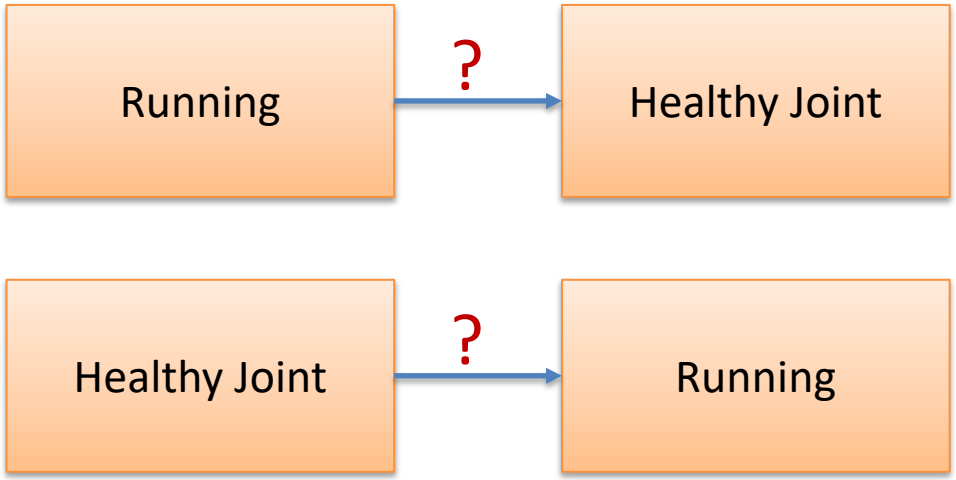
# Reverse Causality



joint_health = $b_0$ + $b_1$*runner +$b_2$*age + $b_3$*gender + ...

If $b_1$ is positive and significant...

- Runners have healthier joints. ✓
- Running improves joint health. ✗

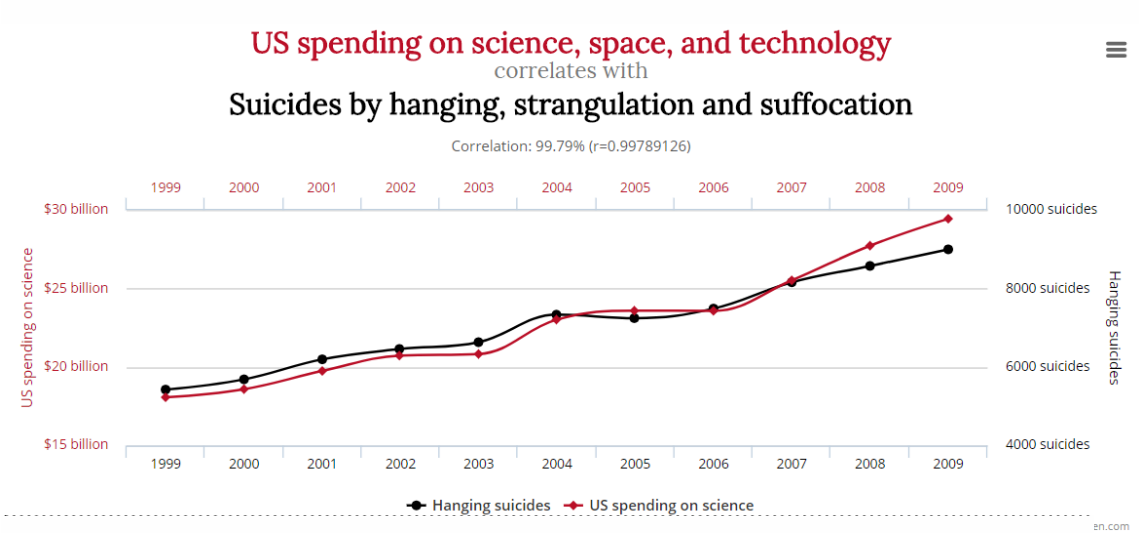


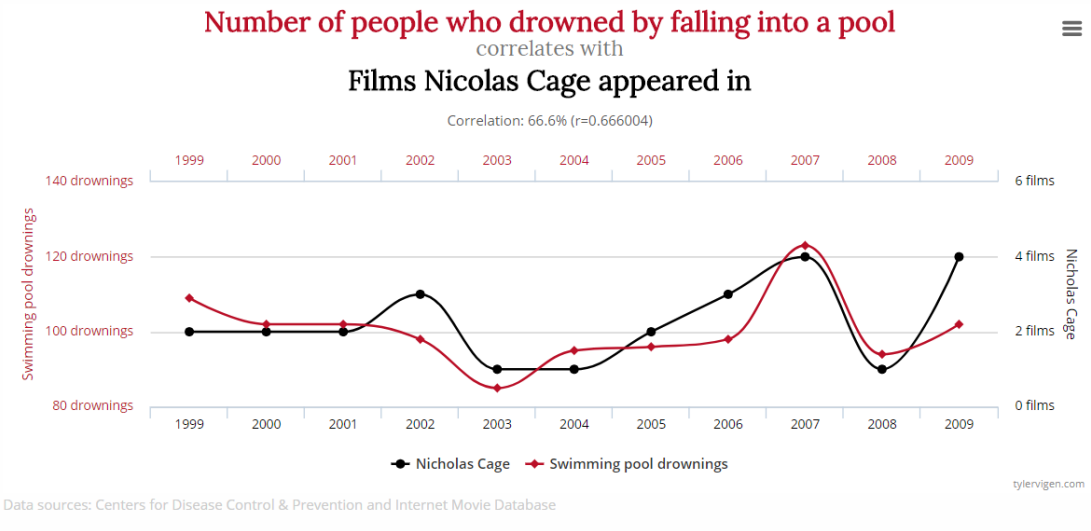
# Self-selection Bias

crime_rate = $b_0$ + $b_1$*green_space

- Individuals can select themselves into a group

# Spurious Regressions with Time-Series Data




Source: https://www.tylervigen.com/spurious-correlations

STEVENS
INSTITUTE of TECHNOLOGY
School of Business

# Thank you!

Prof. Feng Mai
School of Business

For academic use only.