Business Analytics

# Session 3a. Logistic Regression

Renyu (Philip) Zhang

New York University Shanghai

Spring 2018

# Warm-up Exercises

- When can we say that a linear regression model has good predictive power?
  - It should perform well in the out-of-sample test.

- What is the value of including cross terms in a linear regression?
  - Cross terms can capture that the value of one covariate affects the slope of another.

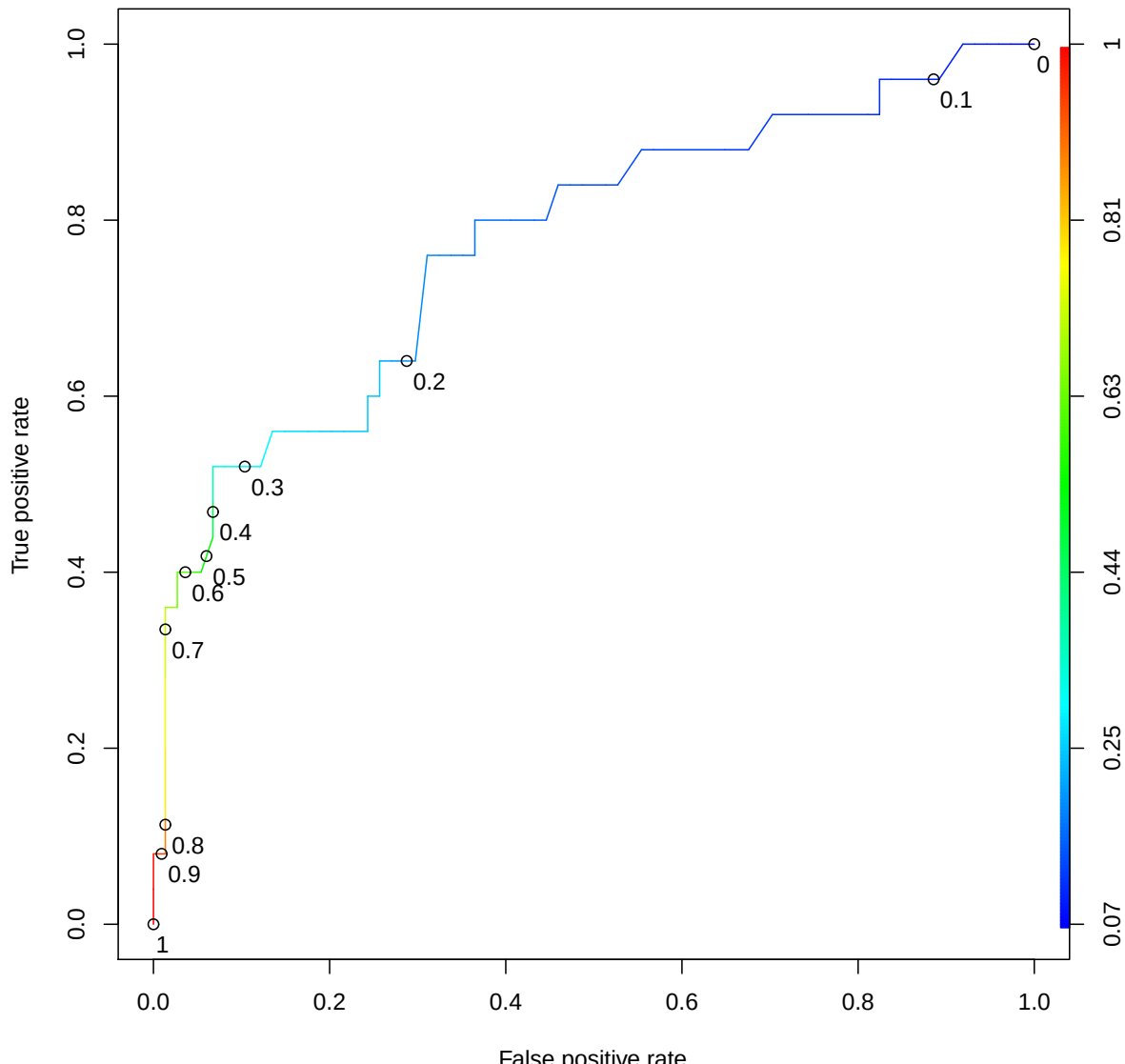
- Our logistic regression model predicts that

$$\mathbb{P}[\mathit{PoorCare} = 1|X] = \frac{\exp(-2.65 + 0.08\mathit{OfficeVisits} + 0.08\mathit{Narcostics})}{1 + \exp(-2.65 + 0.08\mathit{OfficeVisits} + 0.08\mathit{Narcostics})}.$$

  How do we interpret the coefficient of *OfficeVisits*, 0.08?
  - One more office visit is associated with a 0.08 change in the log odds of poor healthcare quality.

# Receiver Operator Characteristic (ROC) Curve

- True positive rate (Sensitivity) on $y-$axis.
- False positive rate (1-specificity) on $x-$axis.
- Best $t$ value trades off the cost of failing to detect positives and the cost of raising false alarms.



# Area Under the ROC Curve (AUC)

- Another measure of model accuracy: Area Under the ROC Curve (AUC).
  - Given a random positive and a random negative outcome, proportion of the time you predict which is which is correct.
  - Less affected by sample (im)balance than accuracy.
  - If the sample is extremely imbalanced, the precision-recall area under the curve is more informative than the ROC-AUC.

- AUC=1 means perfect prediction, AUC=0.5 means pure guessing. In our model, ROC-AUC=0.775.

# Out-of-Sample Accuracy Measures

- Make predictions on a testing data set to compute the out-sample accuracy metrics.

# Out-of-Sample Accuracy Measures

- Make predictions on a testing data set to compute the out-sample accuracy metrics.

- Setting $t = 0.3$, we get the following confusion matrix:

| | Predicted Good Care | Predicted Poor Care |
|---|---|---|
| Actual Good Care | TN=19 | FP=5 |
| Actual Poor Care | FN=2 | TP=6 |

- Overall Accuracy=0.781, Sensitivity=0.750, Specificity=0.792

- AUC=0.799

# Remarks

- The expert-trained analytics model can accurately identify patients receiving low-quality care.
- In practice, the probability outcomes can be used to prioritize patients for interventions.
- Experts are great with a small claim sample, but analytics models are scalable.
- Analytics models can integrate assessments of many experts to achieve unbiased and unemotional predictions.

# More about Logistic Regression and Beyond

# Do you Believe the Model?

- An inherent tension for working with models.
  - The model implies a probabilistic description of the process generated the data (a.k.a. data generating process).
  - Logistic regression seems to be created for technical convenience. (Is the population really logistic?)

- "All models are wrong, but some are more useful than others." — *George E.P. Box*

# Latent Variable Interpretation

- For sample point $i$, define a variable $Z_i$ which follows the distribution: $\mathbb{P}(Z_i \leq z) = \frac{\exp(z)}{1+\exp(z)}$.

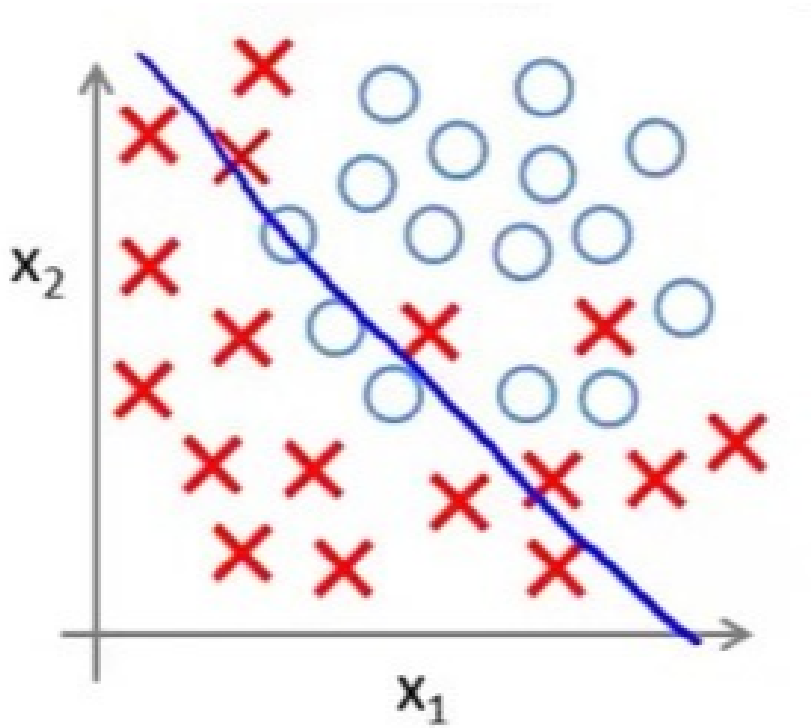
- $Y_i = 1$ if $\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \geq Z_i$; otherwise, $Y_i = 0$.

$$\mathbb{P}(Y_i = 1|X_i) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})} = 1 - \mathbb{P}(Y_i = 0|X_i)$$

- Example: Customers' purchasing choice.

# Logistic Regression as a Linear Classifier

- We predict $Y_i = 1$ if $\hat{\mathbb{P}}(Y_i = 1|X_i) \geq t$, which is equivalent to predicting $Y_i = 1$ if $\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ij} \geq t'$, where $t = \frac{\exp(t')}{1+\exp(t')}$.



# Classification

- Classification problems: Spam filtering, medical diagnostics, admitting new students, etc.

- For a data set $\mathcal{D} = \{Y_i \in \{0,1\}, X_{ij} : 1 \leq i \leq n, 1 \leq j \leq p\}$, find a classifier $\hat{f}(\cdot) \in \{0,1\}$ (a.k.a. a model).

- The model should be evaluated with some metrics of its prediction error under the distribution/data generating process that generates the data.
  - 0-1 loss: $\mathbb{P}[Y \neq \hat{f}(X)|X]$ (the probability is calculated under the "true" data generating process.)
  - FP, FN, OA, ROC, AUC: All these performance measures apply to the general classification problem.

# Building a Classifier

- If we want to minimize 0-1 loss $\mathbb{P}[Y \neq \hat{f}(X)|X]$, we need to use the (naïve) Bayes classifier:
  - $\mathbb{P}[Y = 1|X] = p$, the 0-1 loss for $\hat{f}(X) = 1$ is $p$ and the 0-1 loss for $\hat{f}(X) = 0$ is $1 - p$.
  - If $p > 0.5$, $\hat{f}(X) = 1$; if $p < 0.5$, $\hat{f}(X) = 0$.

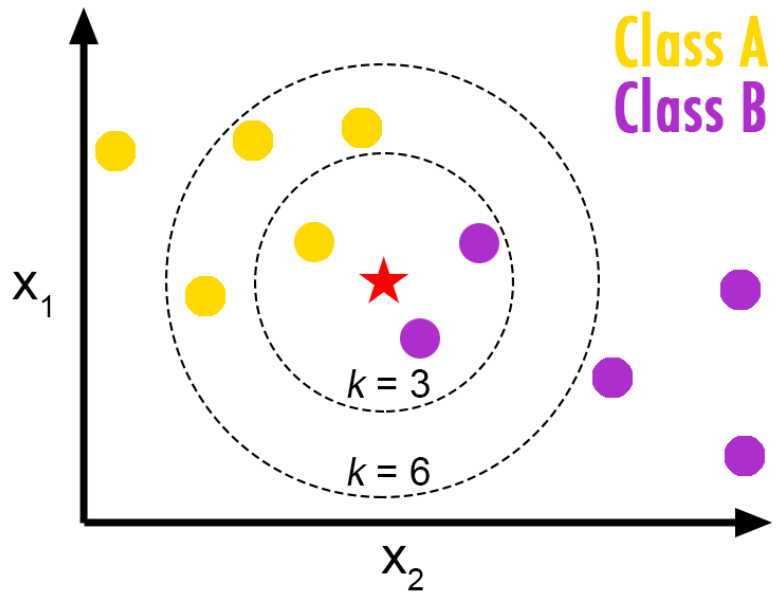
- In a lot of cases, we do not know the underlying population distribution, so it is *impossible* to compute $\mathbb{P}(Y = 1|X)$.
  - Approximation scheme: $k-$nearest neighbors.

# $k-$Nearest Neighbors

- Given a covariate vector $X$, find the $k-$nearest neighbors (usually in Euclidean distance), take the majority vote to determine the classification.

- Intuition: Approximates the Bayes classifier by local averaging (closer covariates will be associated with closer outcomes).

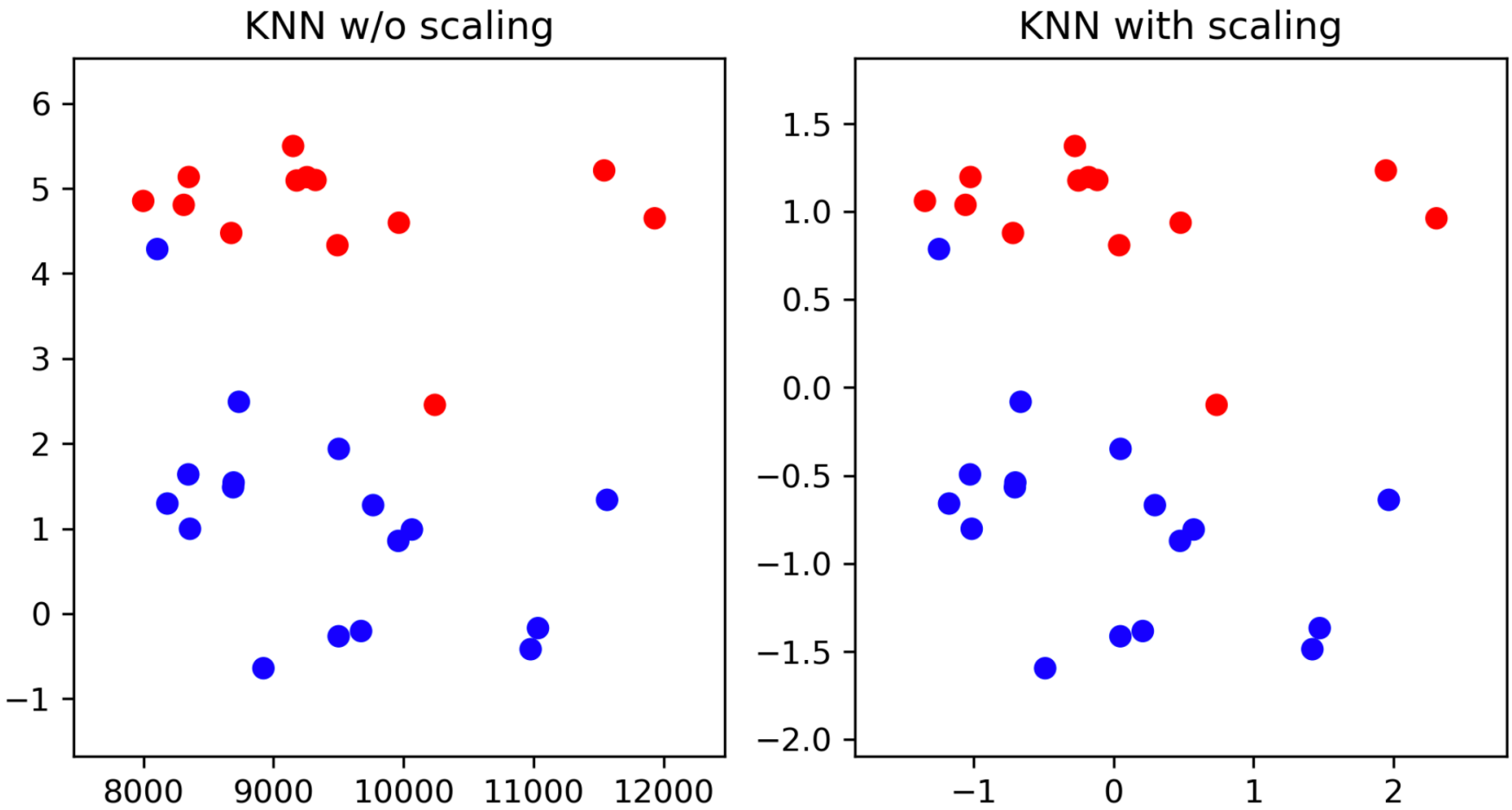
- Parameter $k$: Captures model complexity.
  - The smaller the $k$, the more complex the model.



# Remarks about $k-$Nearest Neighbors

- $k-$NN is a simple yet effective classifier if *distance reliably reflects similarities*.
- $k-$NN is very accurate when the (training data set) sample size $n$ is large, but it is also very slow in this case.
  - Need to calculate the between each testing data and each training data.
- Scale the covariates by standardization.

# $k$–Nearest Neighbors with and without Scaling



# $k$–Nearest Neighbors with and without Scaling