Classification

SLM003 06/08/2018

References:

`ISL04` James G., Witten D., Hastie T., Tibshirani R. (2013) **Classification**. In: *An Introduction to Statistical Learning*. Springer Texts in Statistics, vol 103. Springer, New York, NY. doi: https://doi.org/10.1007/978-1-4614-7138-7_4 (https://doi.org/10.1007/978-1-4614-7138-7_4)

`ESL04` Hastie T., Tibshirani R., Friedman J. (2009) **Linear Methods for Classification**. In: *The Elements of Statistical Learning* (2nd ed.). Springer Series in Statistics. Springer, New York, NY. doi: https://doi.org/10.1007/978-0-387-84858-7_4 (https://doi.org/10.1007/978-0-387-84858-7_4)

Outline

1. Logistic regression
2. Discriminant analysis
    A. Linear discriminant analysis (LDA)
    B. Quadratic discriminant analysis (QDA)

Objectives

- Understand the principles behind the methods
- Develop intuition of the mathematical formulation

What is ”classification”?

- **Supervised learning**: use inputs to predict output
- Classification predicts ***qualitative*** (a.k.a. *categorical*, *discrete*) outputs
- Input: ***predictors*** (a.k.a. *features*, *independent variables*, X) -- quantitative and/or qualitative
- Output: ***response*** (a.k.a. *target*, *dependent variable*, y)
    - which may be refered to as different *response levels*, *targets*, ***classes***, *categories*

Logistic regression

Goal: Describe predictor-response relationship in the *training data* using the **logistic model**. Make prediction using this model.

The *logistic function* (a.k.a. *sigmoid curve*) is defined as:

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)} \tag{4.6}$$

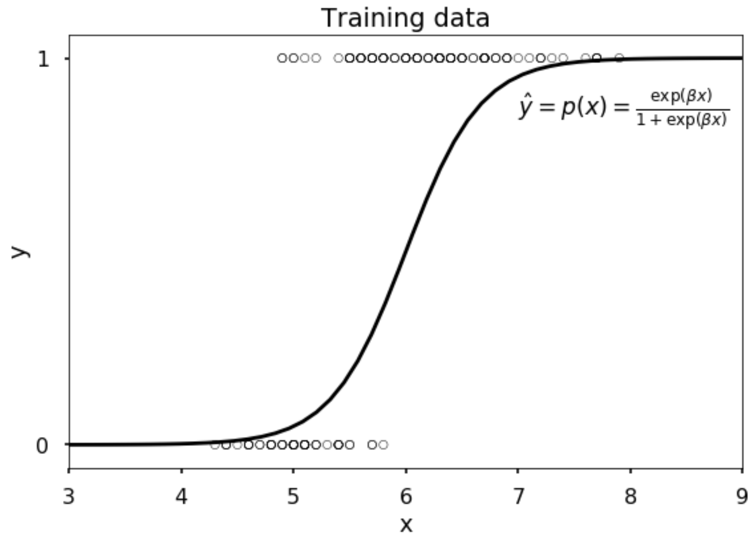
$p(X)$: predicted response | X_i: predictors | β_i: parameters of the model

How to fit the model, i.e. how to determine the appropriate β_i?

Fitting a logistic model using "maximum likelihood"

Focusing on binary response ($k = 2$)

In [5]:
```
interactive(fig_maxlikelihood, page=(0, 5))
```

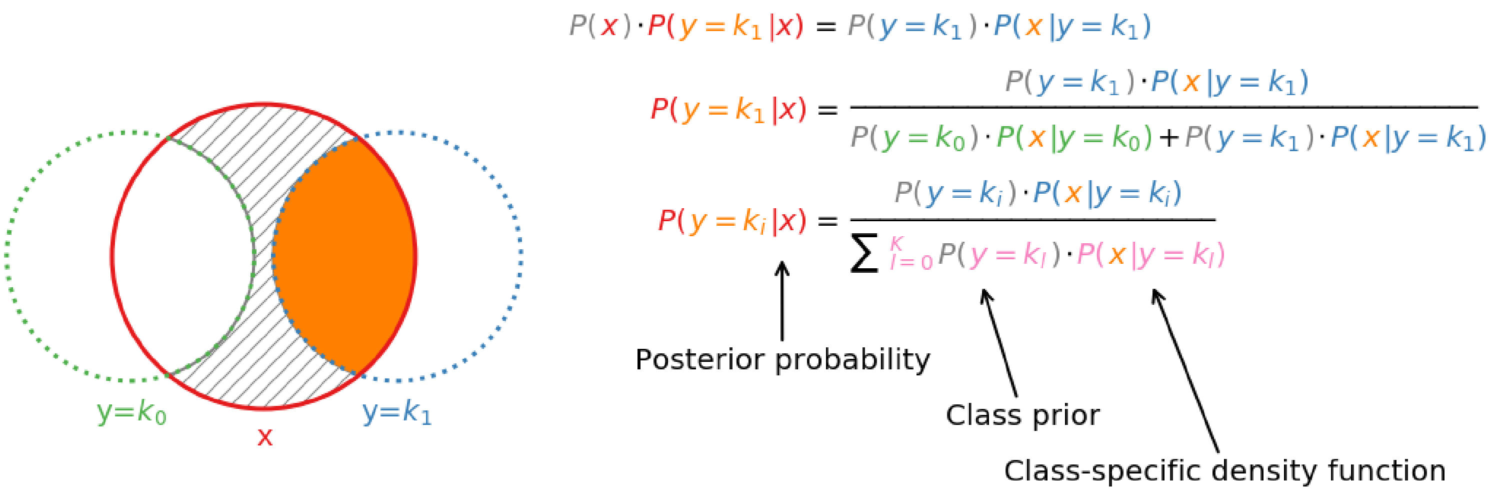


Discriminant analysis

Goal: Assign data to the most probable class based on **distribution statistics** derived from *training data* and/or prior knowledge

Bayes' Theorem

In [7]:
```
interactive(fig_bayes, page=(0,9))
```



Given predictors x, we can determine the probability that the observation belongs to each response class $y = k_i$, if we know the probability of observing each class (*prior*), and the predictor distribution within each class (*density function*).

Classification using discriminant analysis

For classification, we do not need to know the posterior $P(y = k_i|x)$, we need only to know which class k_i has the highest posterior, i.e. $\underset{k_i}{\operatorname{argmax}} P(y = k_i|x)$

Based on assumptions about the density function $P(x|y = k_l)$, we can define a *discriminant function* $\delta(x)$, such that:

$$\underset{k}{\operatorname{argmax}}\, \delta_k(x) = \underset{k}{\operatorname{argmax}}\, P(y = k|x)$$

Assume *Gaussian* (a.k.a. *normal*) density function,

- With **common** *predictor covariance* Σ shared by all classes, we can derive a *linear* discriminant (LDA):
$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} {\mu_k}^T \Sigma^{-1} \mu_k + \log \pi_k \qquad (4.19)$$
- With **class-specific** *predictor covariance* Σ_k, we get a *quadratic* discriminant (QDA):
$$\delta_k(x) = -\frac{1}{2} x^T \Sigma_k^{-1} x + x^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k - \frac{1}{2} \log |\Sigma_k|$$

Errors are not born equal: Thresholding binary classification

As discussed so far, we threshold the predicted probability (both for logistic regression and for discriminant analysis) at 0.5, without distinguishing between different types of errors.

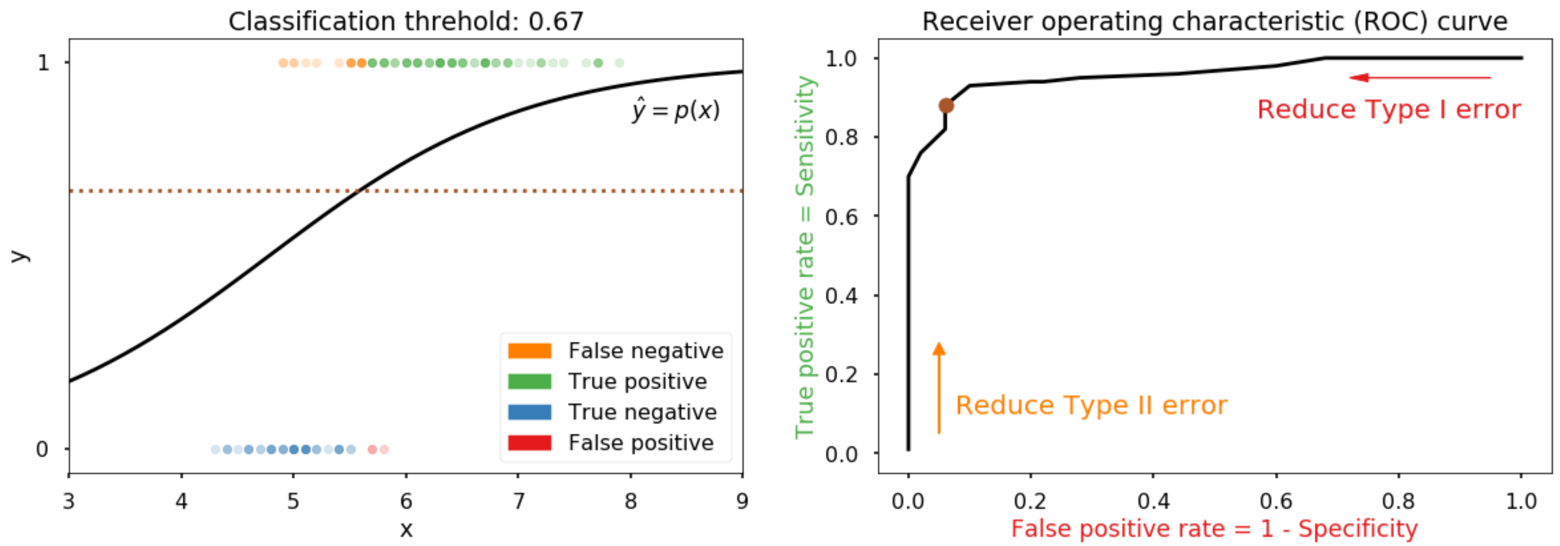
	Pos	Neg
Predict pos	True pos	False pos (Type I error)
Predict neg	False neg (Type II error)	True neg

$$\text{Sensitivity, Recall} = \frac{\text{True pos}}{\text{Pos}} \qquad \text{Specificity} = 1 - \frac{\text{False pos}}{\text{Neg}} = \frac{\text{True neg}}{\text{Neg}}$$

$$\text{Precision} = \frac{\text{True pos}}{\text{Predict pos}} \qquad \text{Accuracy} = \frac{\text{True}}{\text{Total}}$$

Choosing classification threshold based on the ROC curve

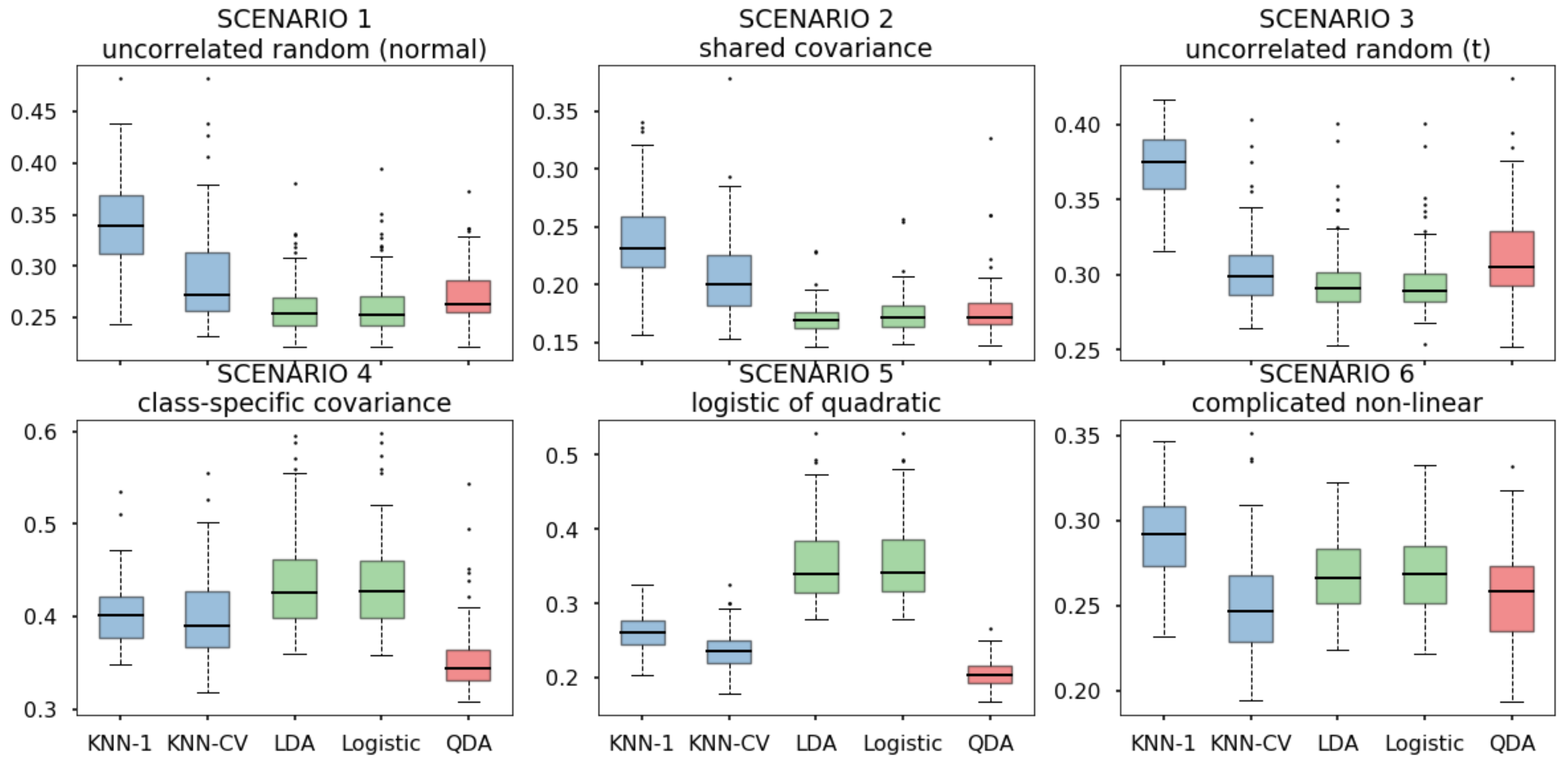
In [9]:
```
interactive(fig_threshold, page=(0,len(thresh_list)-1))
```



Comparison between different classifiers (reproducing Fig 4.10, 4.11)

In [14]:

```
%%time
scores = compare_methods(test_size=2000)
plot_styled(scores)
```




```
CPU times: user 16min 51s, sys: 25.2 s, total: 17min 16s
Wall time: 1min 50s
```

Summary

- Linear decision boundary:
  - Logistic Regression: model each binary decision using a logistic form of *linear* regression, predict the binary response probability
  - LDA: model predictor distributions of each class as *Gaussian*, then based on Bayes' Theorem, compare to see which response is more probable
    - More stable than logistic regression when classes are well-separated
    - If Guassian assumption is valid, more stable than logistic regression when sample size is small
- Non-linear decision boundary:
  - QDA: same as LDA, but allow each class to have *different predictor covariances*
    - More suitable for fitting non-linear decision boundary, but risk overfitting if true boundary is linear