Intégration de données

k Nearest-Neighbors

Ana Karina Fermin

Université Paris Nanterre

M2 Miage APP

http://fermin.perso.math.cnrs.fr/

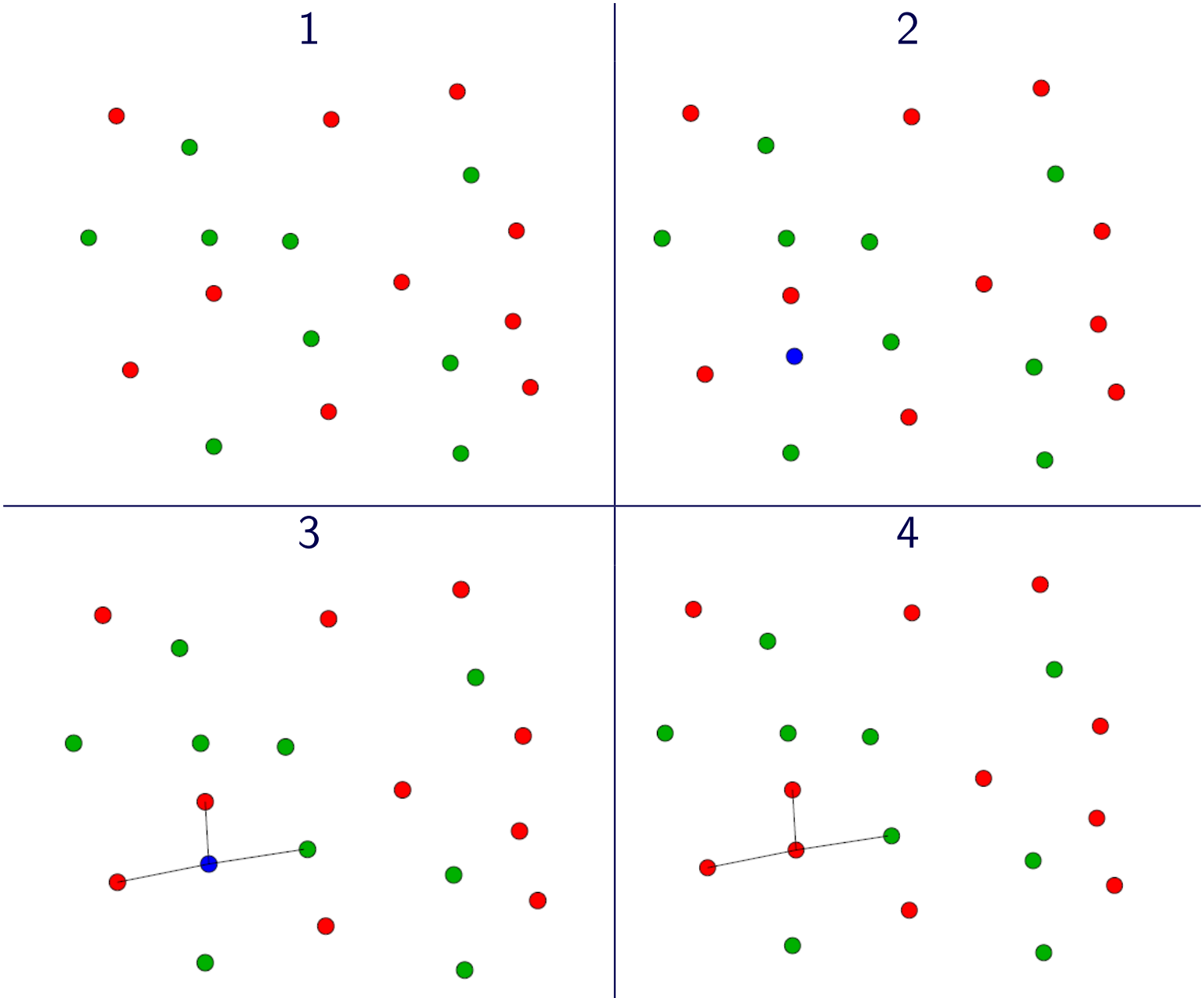
Methods:

1. k Nearest-Neighbors
2. Generative Modeling (Naive Bayes, LDA, QDA)
3. Logistic Modeling
4. Neural Network
5. Tree Based Methods
6. Boosting
7. SVM

1 k Nearest-Neighbors (knn)

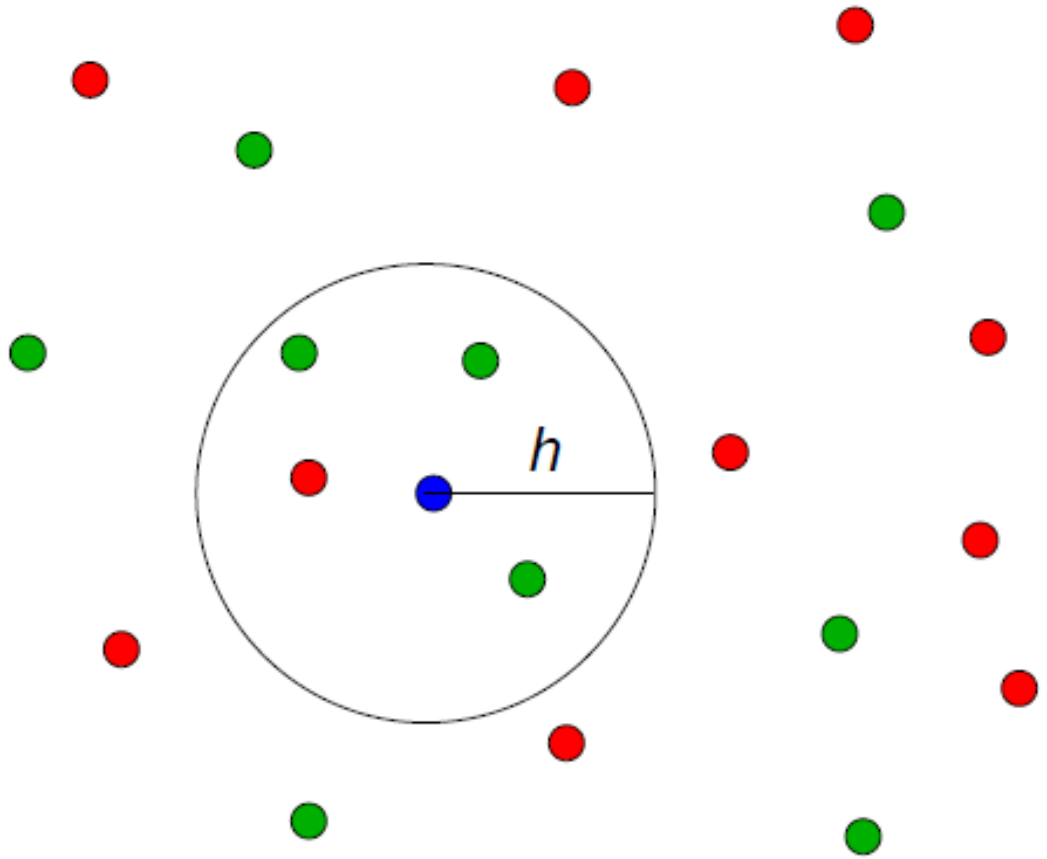
k Nearest-Neighbors (knn)

Example: k Nearest-Neighbors (with $k = 3$)



k Nearest-Neighbors (knn)

Example: k Nearest-Neighbors (with $k = 4$)



- Neighborhood $\mathcal{V}_{\mathbf{x}}$ of $\mathbf{x}$: k closest from $\mathbf{x}$ learning samples.

k-NN as local conditional density estimate

$$\widehat{p}_{+1}(\mathbf{x}) = \frac{\sum_{\mathbf{x}_i \in \mathcal{V}_{\mathbf{x}}} \mathbf{1}_{\{y_i = +1\}}}{k}$$

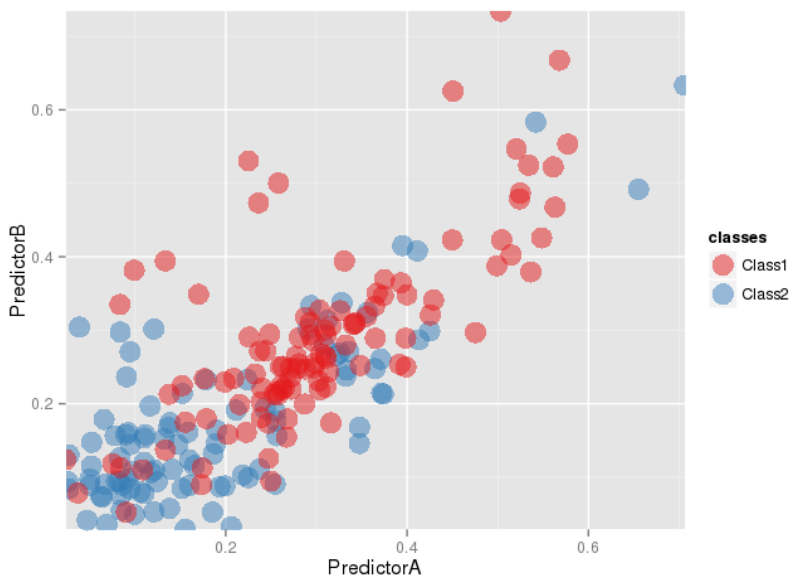
- KNN Classifier:

$$\widehat{f}_{KNN}(\mathbf{x}) = \begin{cases} +1 & \text{if } \widehat{p}_{+1}(\mathbf{x}) \geq \widehat{p}_{-1}(\mathbf{x}) \\ -1 & \text{otherwise} \end{cases}$$

Synthetic Dataset

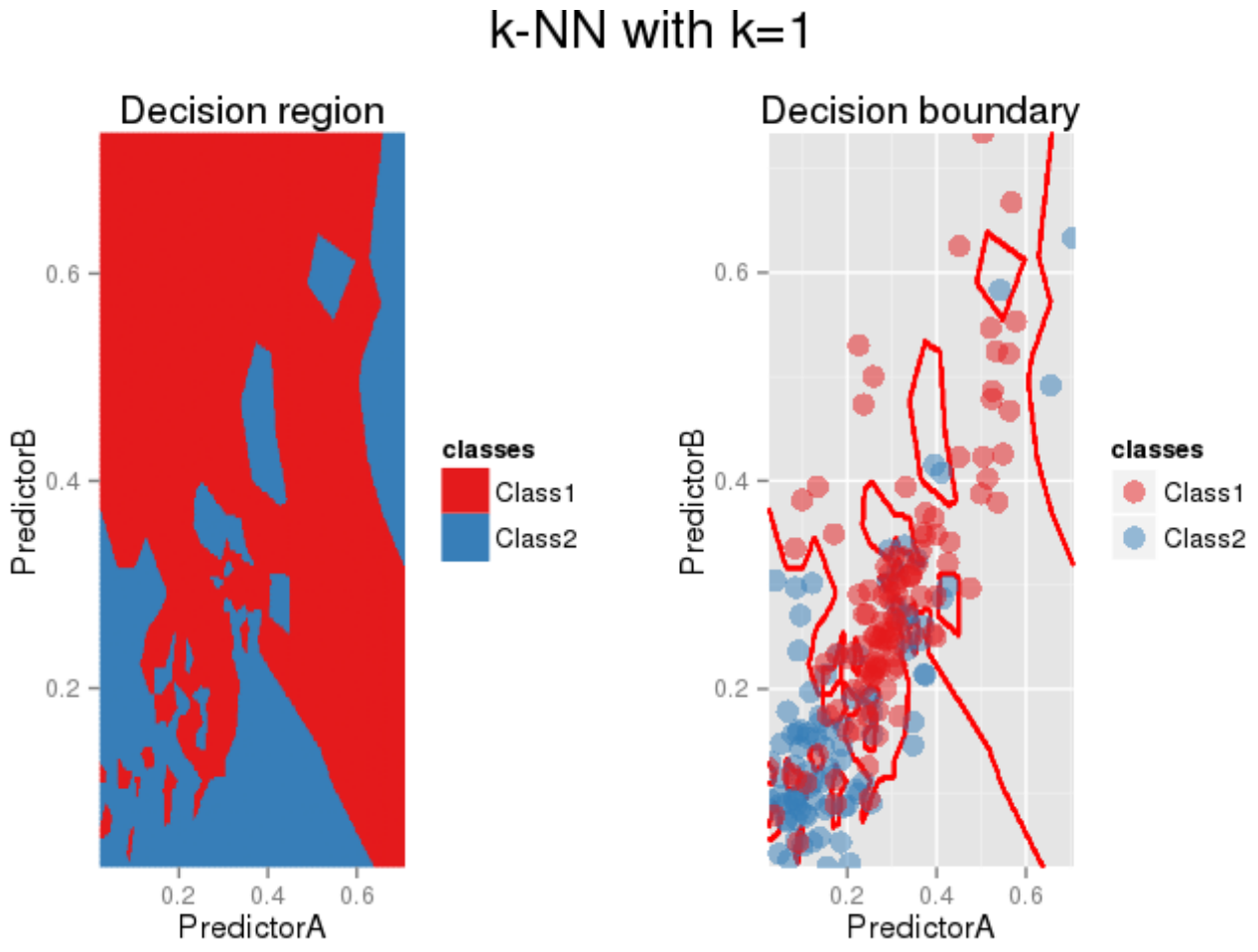
- Two features/covariates.
- Two classes.

- Dataset from *Applied Predictive Modeling*, M. Kuhn and K. Johnson, Springer
- Numerical experiments with **R**.



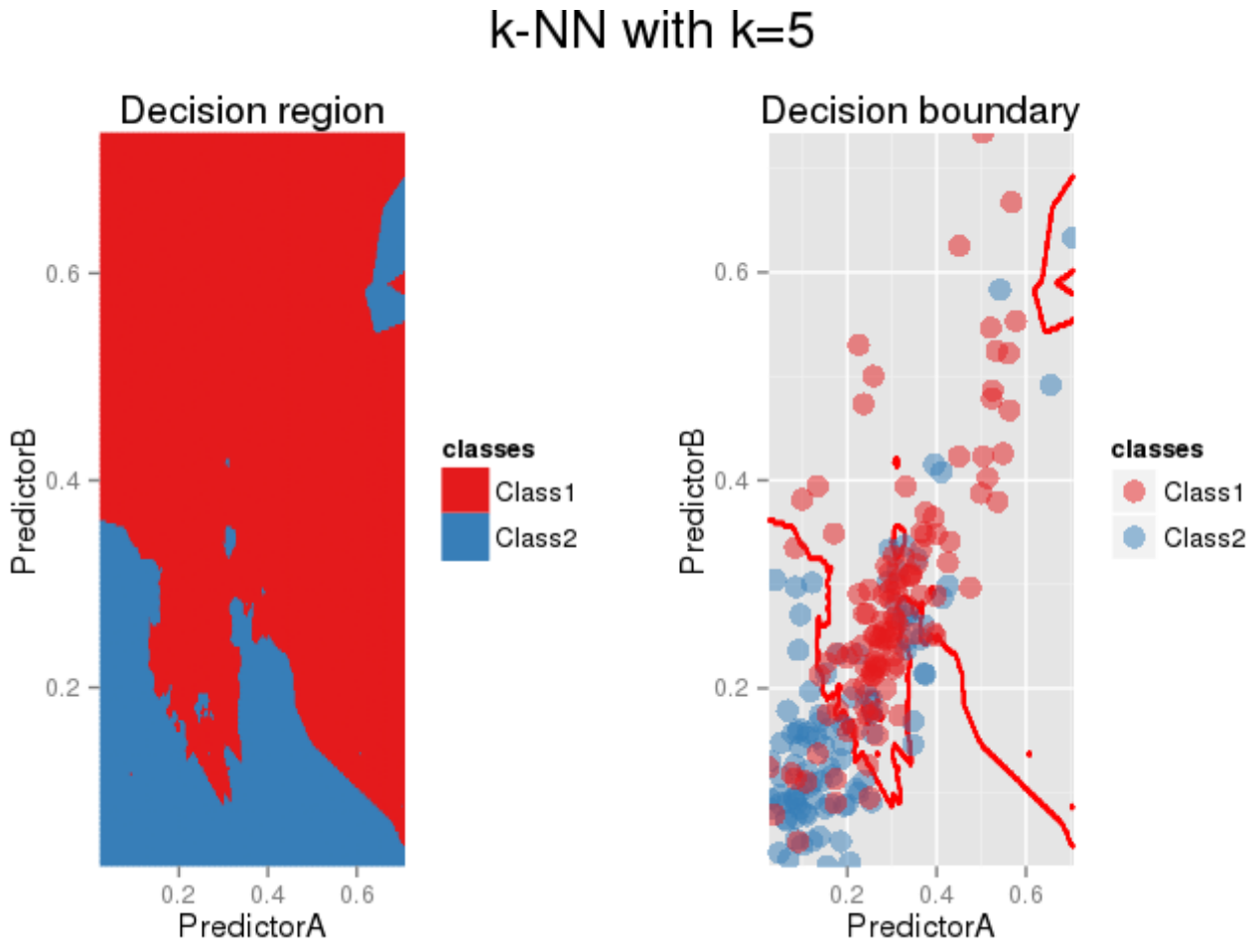
k Nearest-Neighbors (knn)

Example: KNN



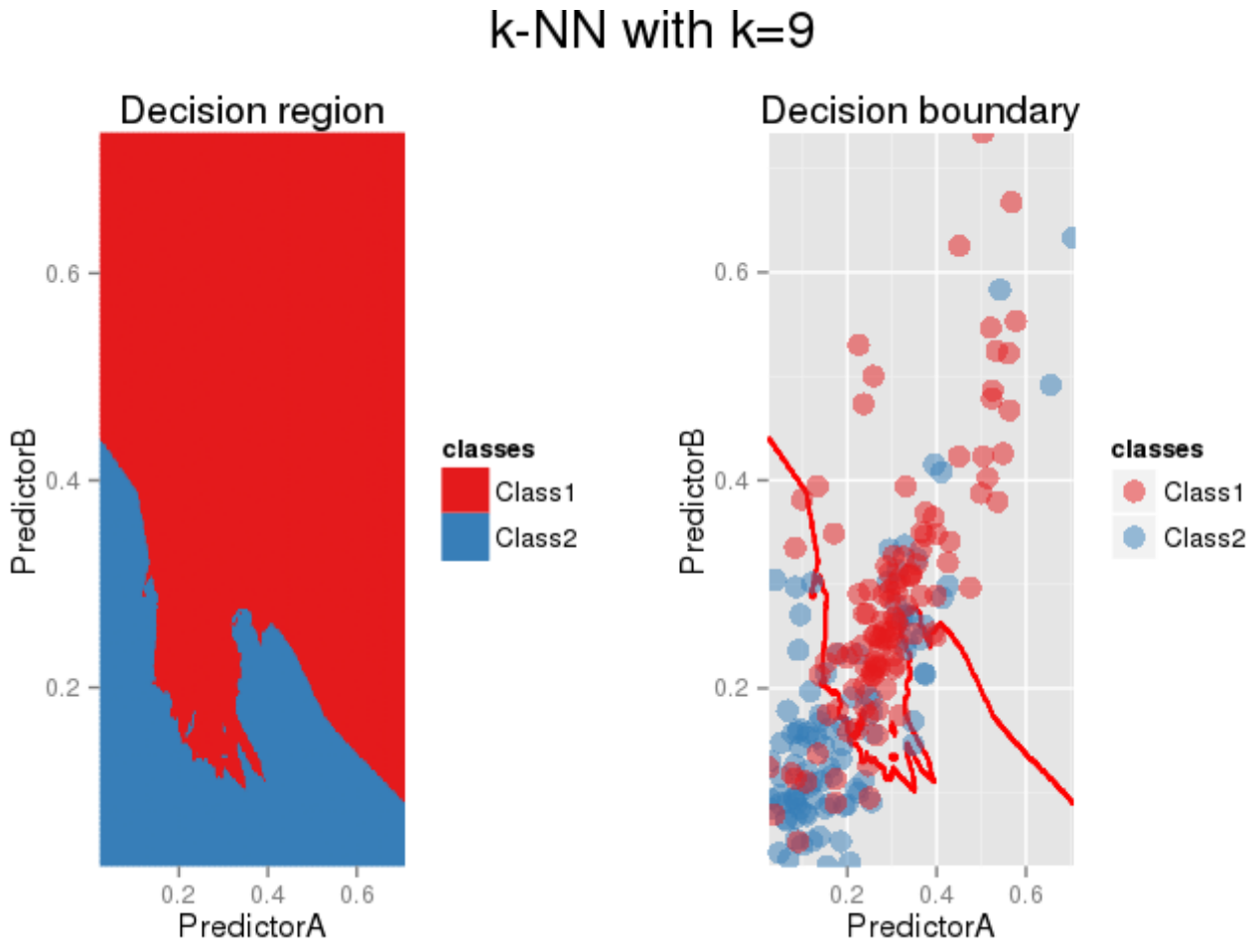
k Nearest-Neighbors (knn)

Example: KNN



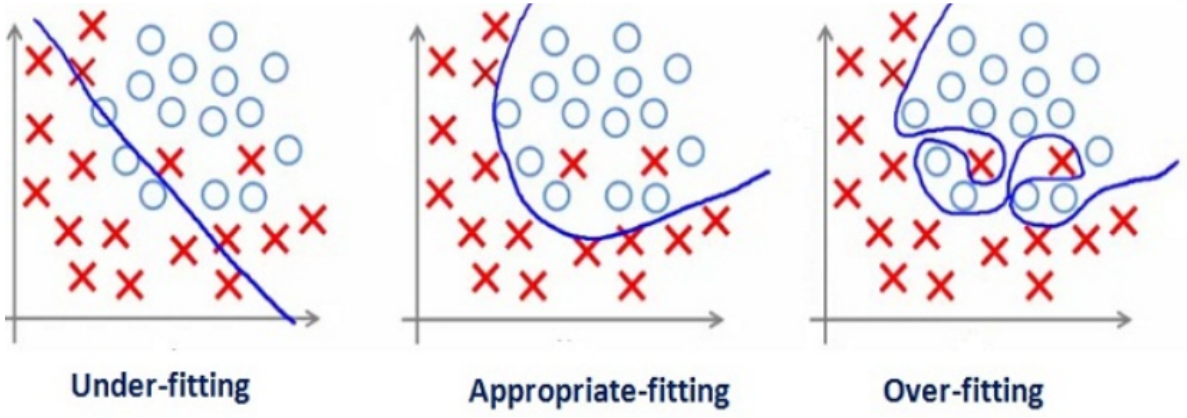
k Nearest-Neighbors (knn)

Example: KNN



k Nearest-Neighbors (knn)

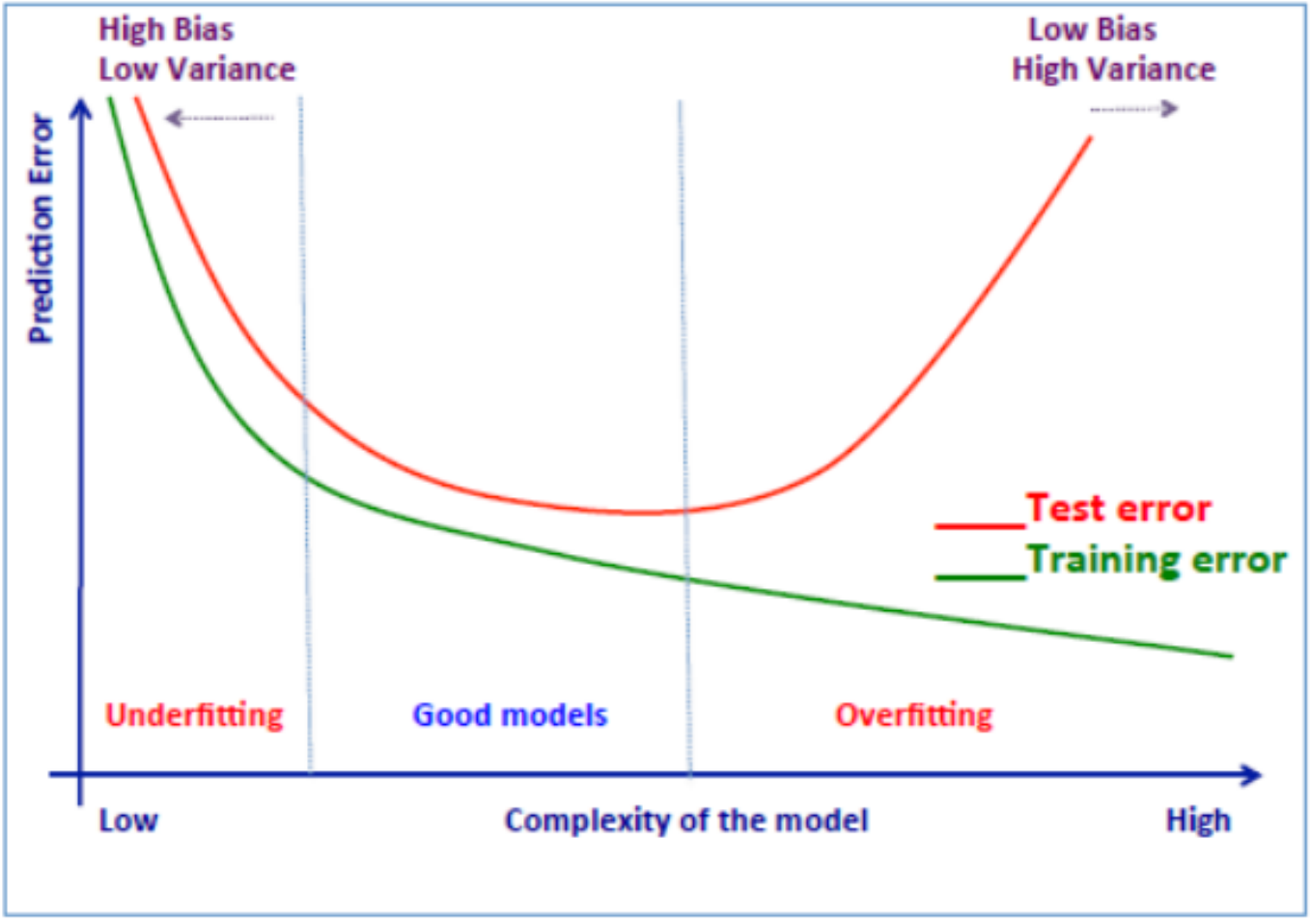
Under-fitting / Over-fitting Issue



- Different behavior for different model complexity
- Under-fit : Low complexity models are easily learned but too simple to explain the truth.
- Over-fit : High complexity models are memorizing the data they have seen and are unable to generalize to unseen examples.

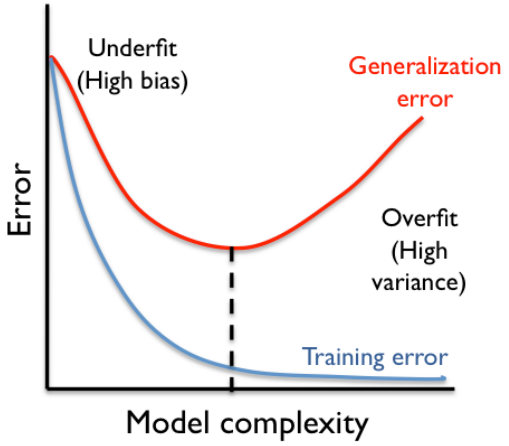
k Nearest-Neighbors (knn)

Under-fitting / Over-fitting Issue



k Nearest-Neighbors (knn)

Over-fitting Issue



Error behaviour

- Learning/training error (error made on the learning/training set) decays when the complexity of the model increases.
- Quite different behavior when the error is computed on new observations (generalization error).

- Overfit for complex models: parameters learned are too specific to the learning set!
- General situation! (Think of polynomial fit...)
- Need to use an other criterion than the training error!

k Nearest-Neighbors (knn)

Example: KNN ($\widehat{k} = 25$ using cross-validation)