Dimensionality Reduction

Instructor: Xiaoqian Wang
10/18/2024

Motivation of Dimensionality Reduction

- Improve learning efficiency
- Improve prediction performance
- Enable better understanding of the learning process with reduced complexity of the learned results

Approaches of Dimensionality Reduction

- Feature Selection
    - Select a subset of features from the original features
- Features Transformation
    - Transform features to replace the original features
- Constructing new features in addition to/instead of the original features
    - General background knowledge (sum or product of features,...)
    - Domain specific background knowledge (clustering of words, parser for text data to get noun phrases,…)

Approaches of Dimensionality Reduction

- Feature Selection
  - Select a subset of features from the original features
- Features Transformation
  - Transform features to replace the original features
- Constructing new features in addition to/instead of the original features
  - General background knowledge (sum or product of features,...)
  - Domain specific background knowledge (clustering of words, parser for text data to get noun phrases,…)

Feature Selection: Example Problem

F_1	F_2	F_3	F_4	F_5	C
0	0	1	0	1	0
0	1	0	0	1	1
1	0	1	0	1	1
1	1	0	0	1	1
0	0	1	1	0	0
0	1	0	1	0	1
1	0	1	1	0	1
1	1	0	1	0	1

- Data set
  - Five Boolean features
  - C = F1 V F2
  - F3 = $\neg$ F2, F5 = $\neg$ F4
- Optimal subset
  - {F1, F2} or {F1, F3}
- Optimization in space of all feature subsets

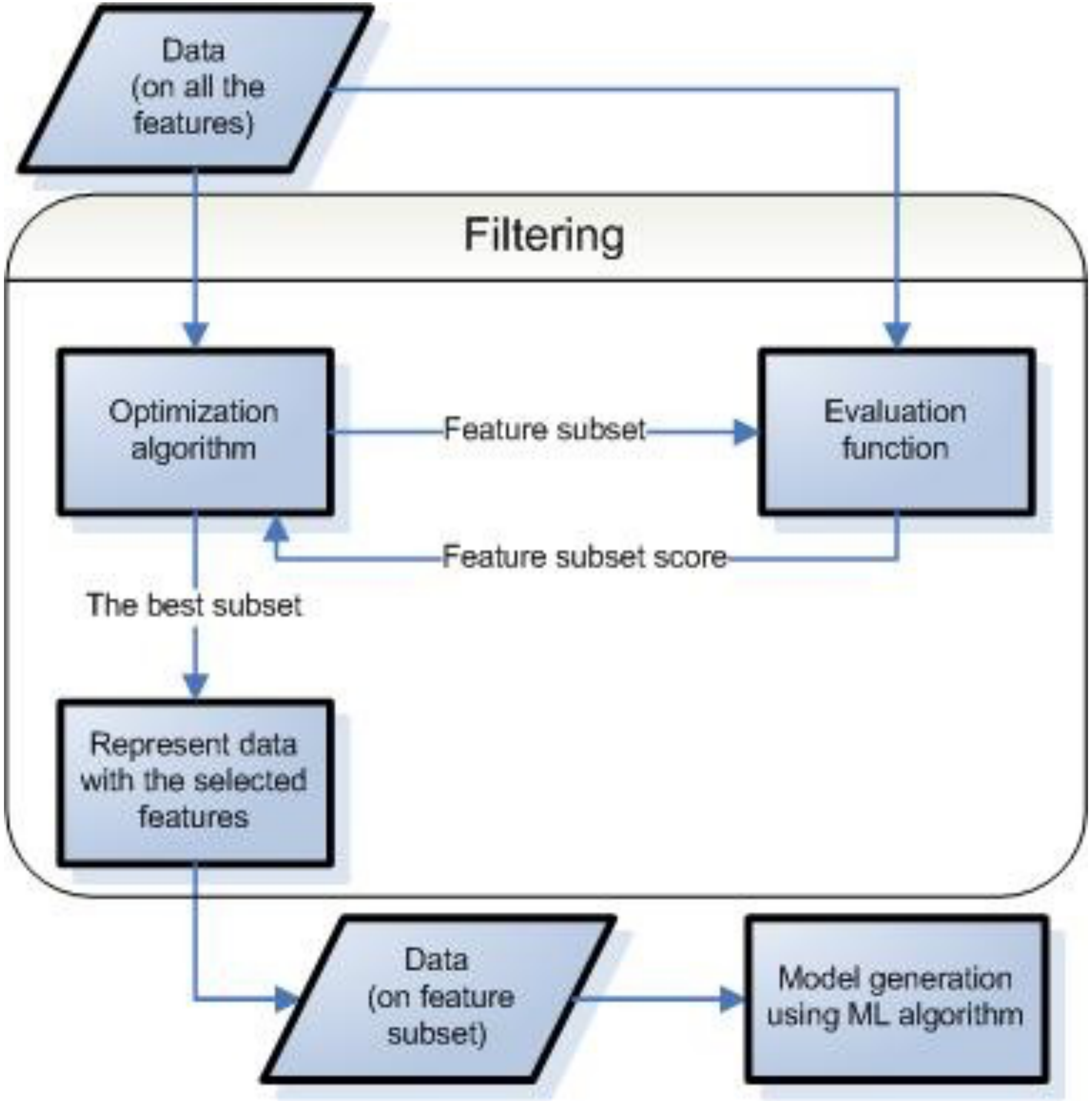
(tutorial on genomics (Yu 2004))

Feature Selection Approaches

- Filter Method
  - Feature selection function independent of the learning algorithm
- Wrapper Method
  - Evaluation using model selection based on the machine learning algorithm
- Embedded Method
  - Feature selection during learning

Filters

- Evaluation independent of ML algorithm

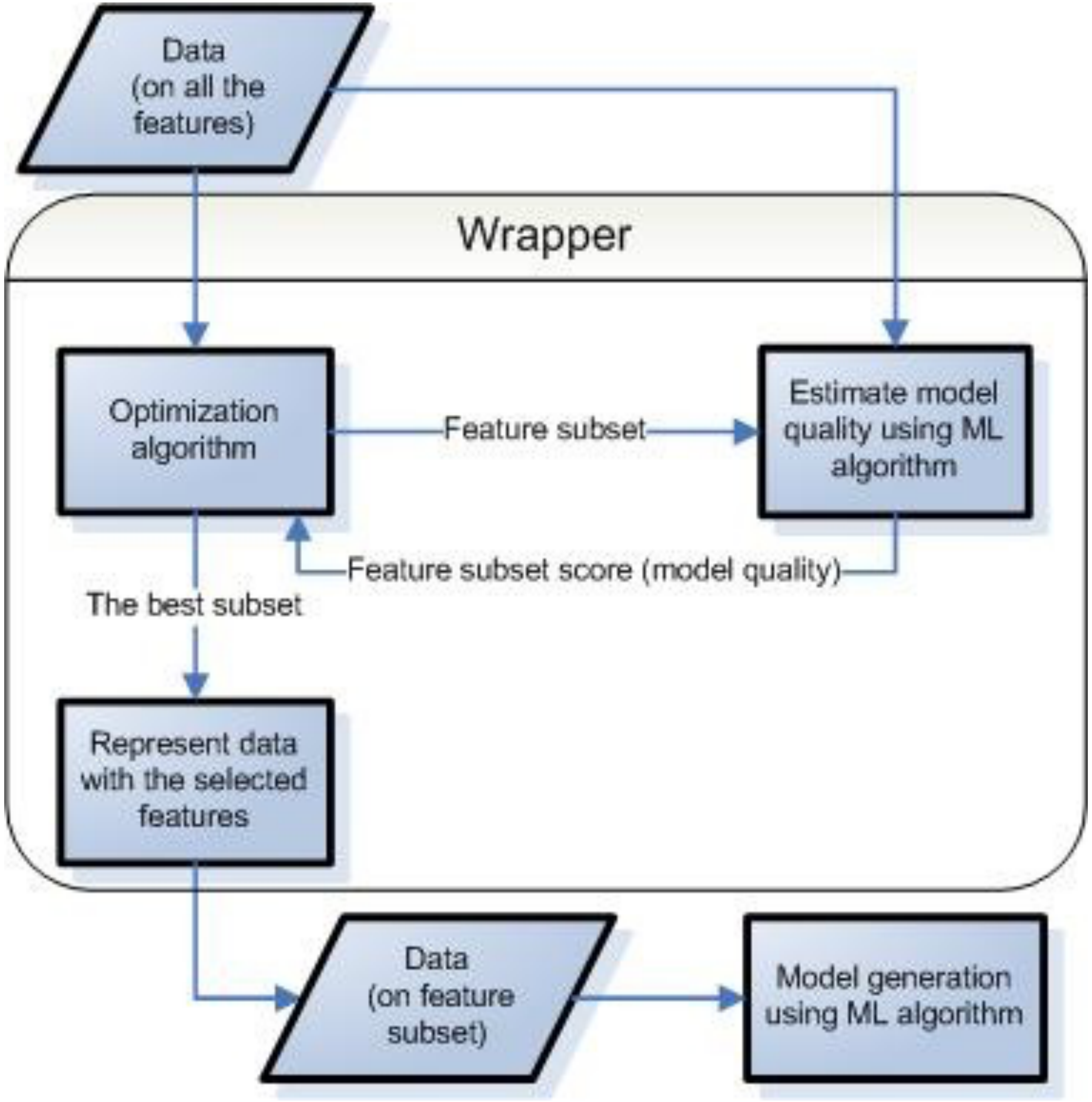


Scoring Individual Feature

- InformationGain: $\sum_{F=W,\overline{W}} P(F) \sum_{C=pos,neg} P(C \mid F) \log P(C \mid F)$
- CrossEntropyTxt: $P(W) \sum_{C=pos,neg} P(C \mid W) \log P(C \mid W)$
- OddsRatio: $\log \frac{P(W \mid pos) \times (1 - P(W \mid neg))}{(1 - P(W \mid pos)) \times P(W \mid neg)}$
- Frequency: $Freq(W)$

Wrappers

- Evaluation uses the same ML algorithm that is used after the feature selection



Wrappers: Drop-Out-One Loss Approach

- Evaluation using Neural Network *(Ye & Sun 2018)*

1) Train a penalized neural network

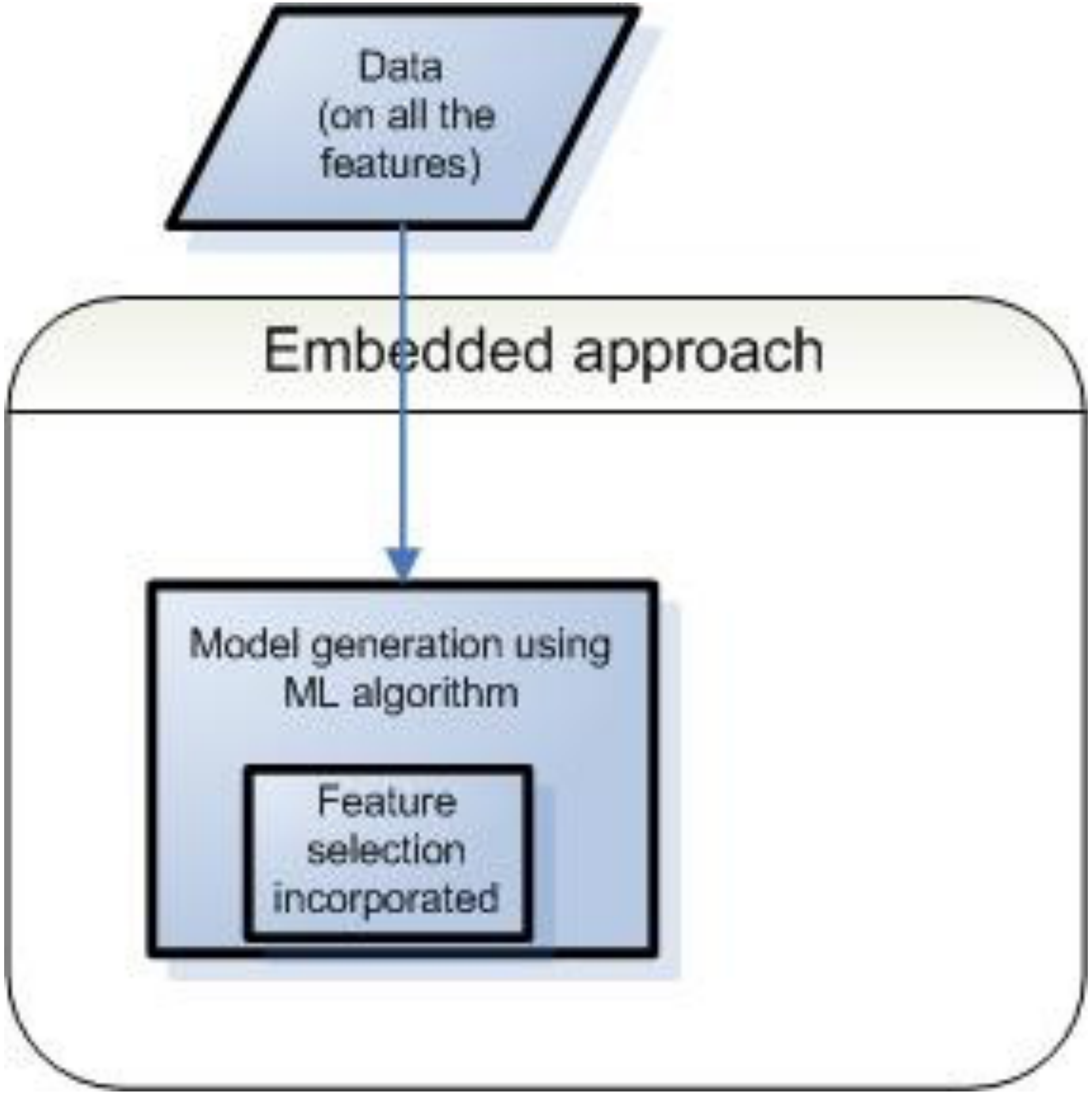
2) Estimate the features by change in loss function

$$\begin{aligned} &\Delta_{\tilde{n}}\mathcal{L}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \\ = \ & \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \left\{ \ell\left(\tilde{y}^{(i)} - f_{\boldsymbol{\eta}_1}(\tilde{\boldsymbol{x}}^{(i)})\right) - \ell\left(\tilde{y}^{(i)} - f_{\boldsymbol{\eta}_2}(\tilde{\boldsymbol{x}}^{(i)})\right)\right\} \end{aligned}$$

- Eliminate individual/ group of feature if the change in loss is smaller than the preset threshold.

Embedded Method

- Feature selection as integral part of model generation



Embedded: with Sparsity Regularization

- Feature selection with sparsity regularization:

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} L_{CE}(y_i, \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x_i}}}) + \lambda \|\mathbf{w}\|_1$$

Embedded: with Sparsity Regularization

- Feature selection with sparsity regularization:

$$\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} L_{CE}(y_i, \frac{1}{1+e^{-\mathbf{w}^\top \mathbf{x_i}}}) + \lambda \|\mathbf{w}\|_1$$

- Feature selection for multi-classes with joint sparsity: $\ell_{2,1}$-norm regularization (Nie & Huang & Cai & Ding 2010)

Approaches of Dimensionality Reduction

- Feature Selection
  - Select a subset of features from the original features
- Features Transformation
  - Transform features to replace the original features
- Constructing new features in addition to/instead of the original features
  - General background knowledge (sum or product of features,...)
  - Domain specific background knowledge (clustering of words, parser for text data to get noun phrases,…)

Principle Component Analysis (PCA)

- Covariance $Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$
- Objective function of PCA

$$\max_{W^\top W = I} tr(W^\top X H X^\top W), \text{where } H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$$

Illustration: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Principle Component Analysis (PCA)

- Discussion of PCA
  - Principle components are orthogonal
  - Sensitive to scale of features
  - Max variance after feature transformation
  - Linear feature transformation

t-SNE (t-distributed Stochastic Neighbor Embedding)

- Nonlinear feature transformation
- Capture local structure of data
- Objective function of t-SNE [Van der Maaten, L., & Hinton, G., 2008]:
  - Pairwise similarity for original data

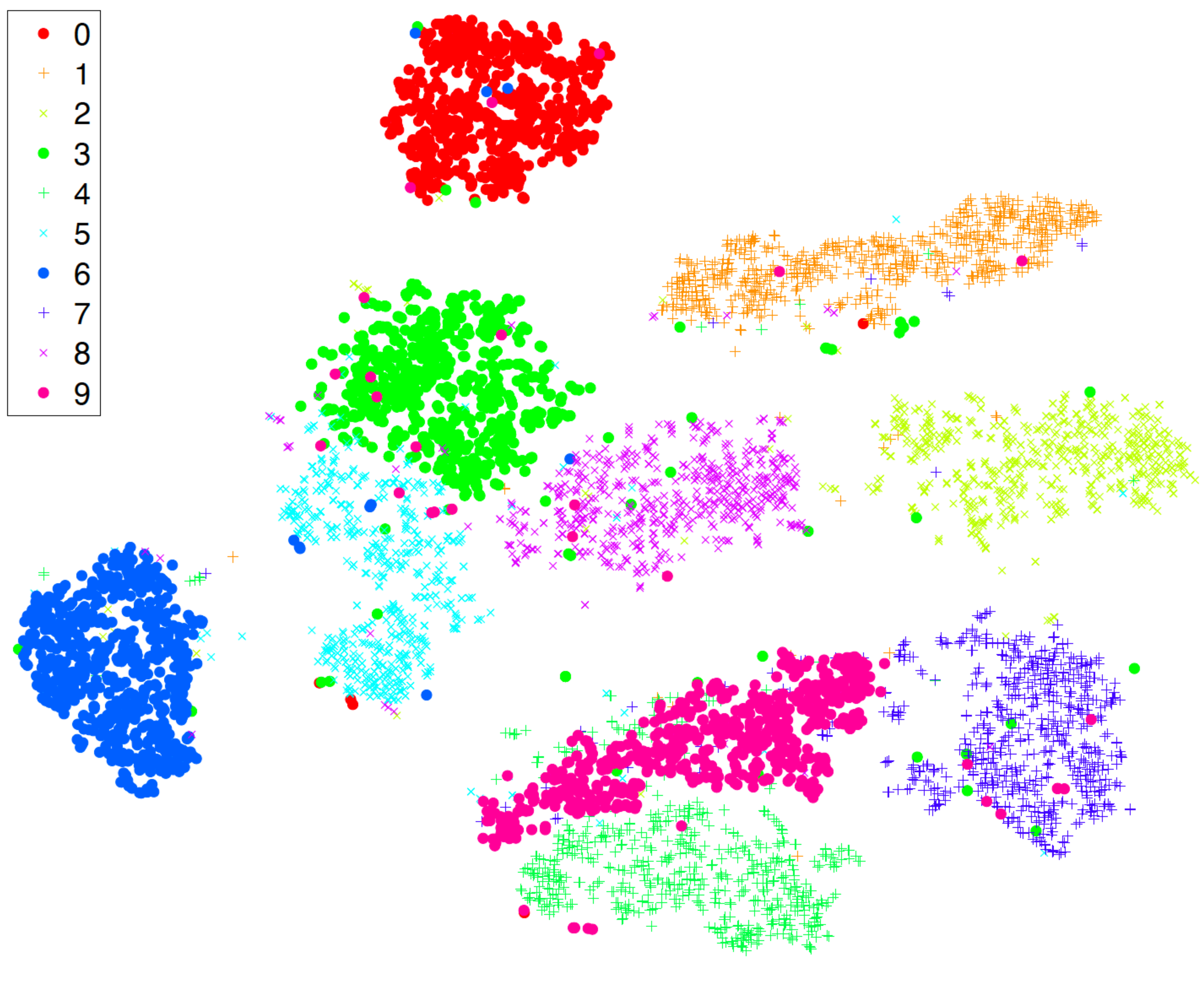
$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)},$$

  - Pairwise similarity for projected data

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}$$

  - KL divergence between distribution P and Q

$$C = \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$



(a) Visualization by t-SNE.

Source: Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

The KL Divergence is Asymmetric (slides from week 2 - probability)

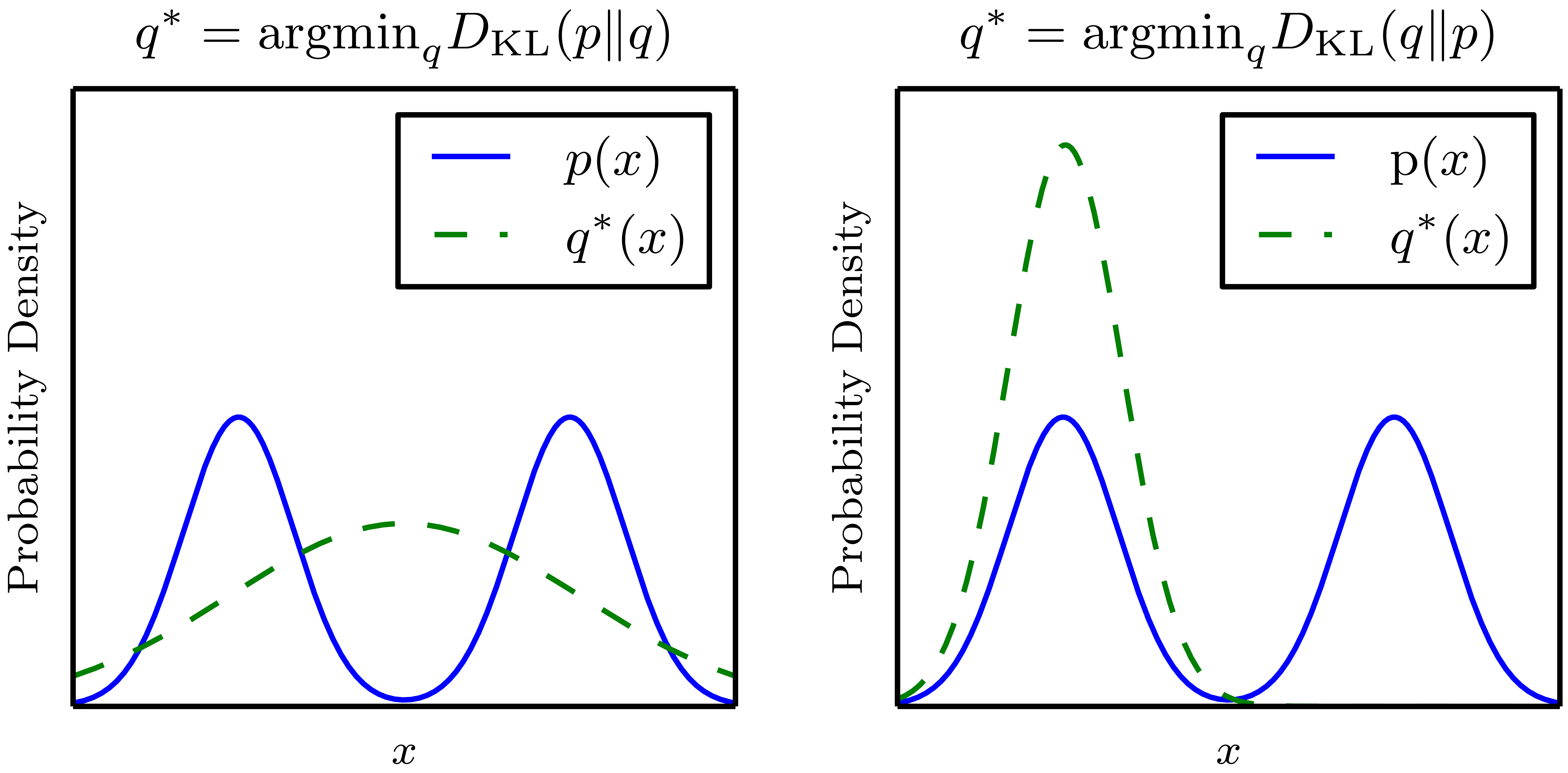


Figure 3.6

(Goodfellow 2016)