Statistical Machine Learning

Lecture 07: Clustering and Evaluation

Kristian Kersting

TU Darmstadt

Summer Term 2020

Today's Objectives

- Make you understand how to find meaningful groups of data points and evaluating the performance of estimators
- Covered Topics:
  - Clustering
  - Bias & Variance
  - Cross-Validation

Outline

1. Clustering

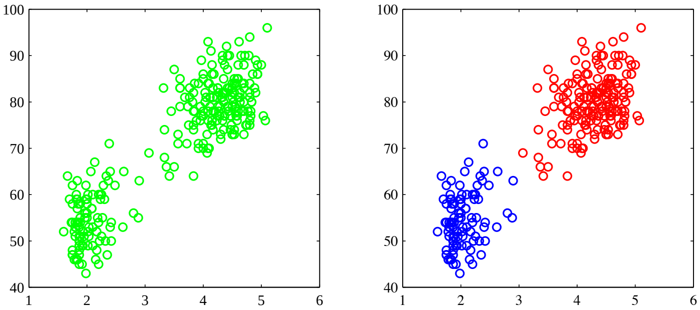
2. Evaluation

3. Wrap-Up

Outline

Clustering

- We introduced mixture models as part of density estimation
- They are also very useful for clustering
  - Divide the feature space into meaningful groups
  - Find the group assignment



Clustering

- Clustering is a type of Unsupervised Learning
- Examples
  - k-Means
  - Mixture models

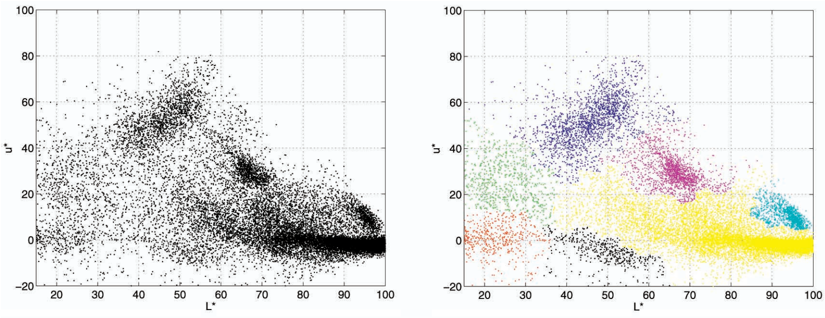
TECHNISCHE UNIVERSITÄT DARMSTADT

Simple Clustering Methods

- Agglomerative Clustering
  - Make each point a separate cluster
  - While the clustering is not satisfactory
    - Merge the two clusters with the smallest inter-cluster distance
- Divisive Clustering
  - Construct a single cluster containing all points
  - While the clustering is not satisfactory
    - Split the cluster that yields the two components with the largest inter-cluster distance

Mean Shift Clustering

- Mean shift is a method for finding modes in a cloud of data points where the points are most dense

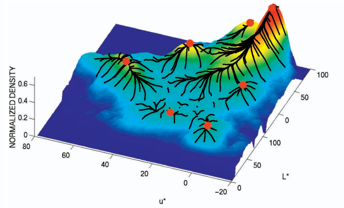


[Comaniciu & Meer, 02]

TECHNISCHE UNIVERSITÄT DARMSTADT

Mean Shift Clustering

- The mean shift procedure tries to find the modes of a kernel density estimate through local search



[Comaniciu & Meer, 02]

- The black lines indicate various search paths starting at different points
- Paths that converge at the same point get assigned the same label

TECHNISCHE UNIVERSITÄT DARMSTADT

Mean Shift Clustering

- Start with kernel density estimate

$$\hat{f}(\mathbf{x}) = \frac{1}{Nh^d} \sum_{i=1}^{N} k\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right)$$

- We can derive the mean shift procedure by taking the gradient of the kernel density estimate
- For details see: D. Comaniciu, P. Meer, *Mean Shift: A Robust Approach toward Feature Space Analysis*, IEEE Trans. Pattern Analysis Machine Intell., Vol. 24, No. 5, 603-619, 2002.

Mean Shift Clustering

TECHNISCHE UNIVERSITÄT DARMSTADT

- Start at a random data point $\mathbf{x}$
- Compute the mean shift vector:

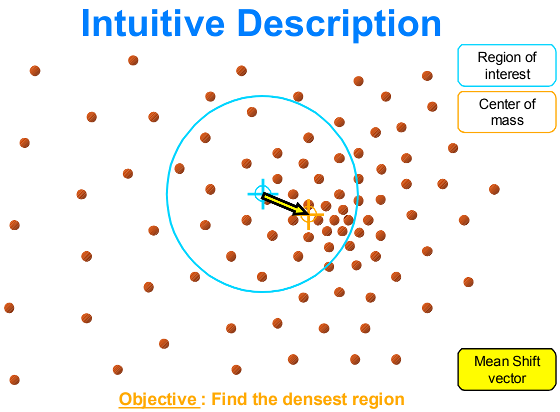
$$m_{h,g}(\mathbf{x}) = \frac{\sum_{i=1}^{N} \mathbf{x}_i g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{N} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x}$$

  - Where $g(y) = -k'(y)$
- Move the current point by the mean shift vector:

$$\mathbf{x} \leftarrow \mathbf{x} + \mathbf{m}_{h,g}(\mathbf{x})$$

- Repeat until convergence

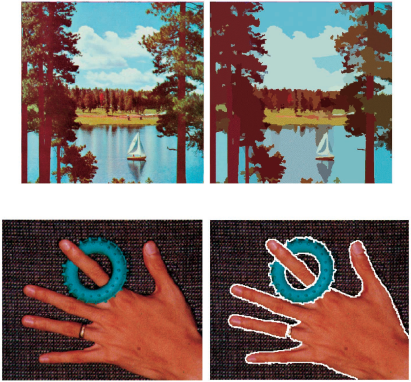
Mean Shift Clustering - Illustration



[Ukrainitz & Sarel]

Segmentation using Clustering

- Clustering of simple image features, e.g. color & pixel position



[Comaniciu & Meer, 02]

TECHNISCHE UNIVERSITÄT DARMSTADT

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Evaluation

- What have we seen so far...
  - Classification using the Bayes classifier

$$p(C_k \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid C_k)\, p(C_k)$$

  - Probability density estimation to estimate the class-conditional densities $p(\boldsymbol{x} \mid C_k)$
- How do we know how well we are carrying out each of these tasks?
- We need a way of performance evaluation
  - For density estimation (or really, parameter estimation)
  - For the classifier as a whole

TECHNISCHE UNIVERSITÄT DARMSTADT

Evaluation

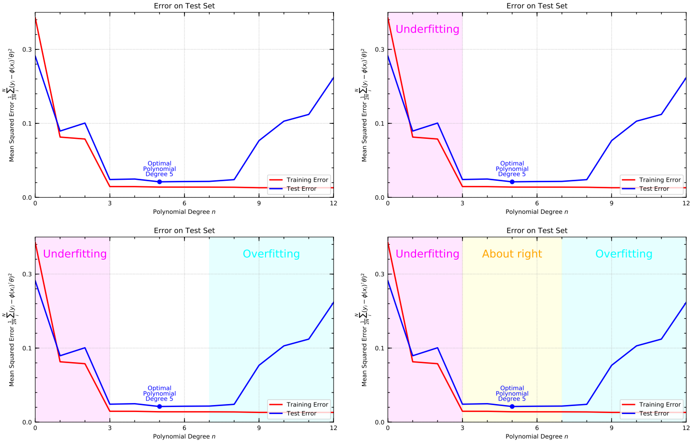
- **Overfitting is everywhere**


Is there a tank in the picture?

[DARPA Neural Network Study (1988-89), AFCEA International Press]

TECHNISCHE UNIVERSITÄT DARMSTADT

Test Error vs Training Error



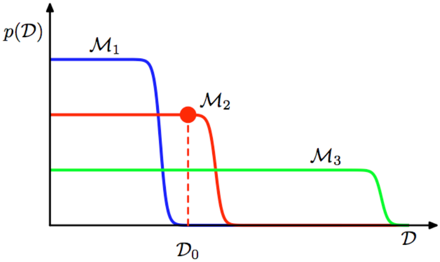
Small training error good model?

⇒**NO! We need to rethink model selection!**

Occam's Razor and Model Selection

William of Ockham
(1285-1347)

- **Model Selection Questions**
  - Number of parameters a.k.a. degree of polynomial n?
  - Is your model class sufficently rich? $\Rightarrow$ Underfitting
  - Too rich? $\Rightarrow$ Overfitting
- **Occams's Razor**:
  - Always choose the simplest model that fits the data
  - Simplest = smallest model complexity!



TECHNISCHE UNIVERSITÄT DARMSTADT

Bias and Variance

- As we saw before, maximum likelihood is just one possible way to estimate a parameter
  - How can we assess how good an estimator is?
- Assume that we have an estimator $\hat{\theta}$ that estimates the parameter θ from the data set $\boldsymbol{X}$
- Bias of an estimator: expected deviation from the true parameter

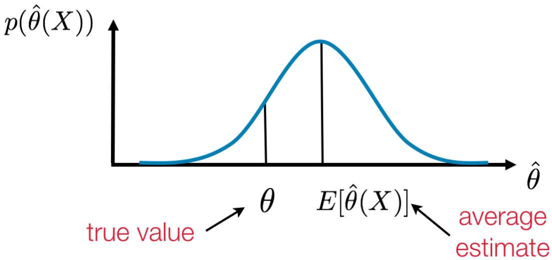
$$\text{bias}\left(\hat{\theta}\right) = \mathbb{E}_{\mathbf{X}}\left[\hat{\theta}\left(\mathbf{X}\right) - \theta\right]$$

- Variance of an estimator: expected squared error between the estimator and the mean estimator

$$\text{var}\left(\hat{\theta}\right) = \mathbb{E}_{\mathbf{X}}\left[\left\{\hat{\theta}\left(\mathbf{X}\right) - \mathbb{E}_{\mathbf{X}}\left[\hat{\theta}\left(\mathbf{X}\right)\right]\right\}^2\right]$$

Bias of an Estimator

- The estimate $\hat{\theta}\left(\mathbf{X}\right)$ is a random variable, because we assumed that $\mathbf{X}$ is a random sample from a true underlying distribution

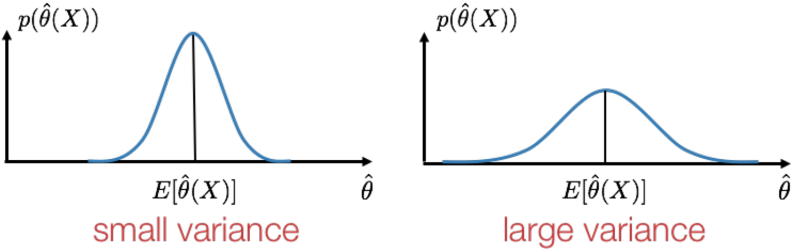


- An estimator is biased if the expected value of the estimator $\mathbb{E}_{\mathbf{X}}\left[\hat{\theta}\left(\mathbf{X}\right)\right]$ differs from the true value of the parameter θ

- Otherwise it is called unbiased, i.e., $\mathbb{E}_{\mathbf{X}}\left[\hat{\theta}\left(\mathbf{X}\right)\right] = \theta$

TECHNISCHE UNIVERSITÄT DARMSTADT

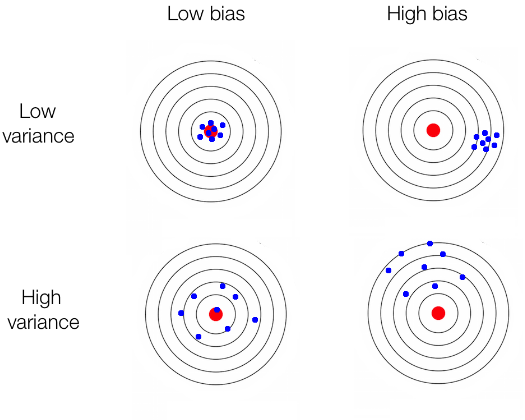
Variance of an Estimator



- Ideally, we want an unbiased estimator with small variance
- In practice, this is not that easy as we will see shortly

TECHNISCHE UNIVERSITÄT DARMSTADT

Bias and Variance



TECHNISCHE UNIVERSITÄT DARMSTADT

I am so BLUE...

- An estimator with
  - Zero bias
  - Minimum variance

  is called a Minimum Variance Unbiased Estimator (MVUE)
- A Minimum Variance Unbiased Estimator which is
  - Linear in the features

  is called a Best Linear Unbiased Estimator (BLUE)

Maximum-Likelihood Estimation (MLE) of a Gaussian

- Remember, the Gaussian has two parameters, the mean μ and the variance σ^2
- Let's compute the bias of the Maximum-Likelihood estimate of the mean of a Gaussian

$$
\begin{aligned}
\hat{\mu}(X) &= \frac{1}{N}\sum_{i=1}^{N} x_i \\
\mathbb{E}_{\mathbf{X}}\left[\hat{\mu}(\mathbf{X}) - \mu\right] &= \mathbb{E}\left[\hat{\mu}(\mathbf{X})\right] - \mathbb{E}[\mu] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] - \mu \\
&= \frac{1}{N}\left(\sum_{i=1}^{N}\mathbb{E}[x_i]\right) - \mu = \frac{1}{N}\left(\sum_{i=1}^{N}\mu\right) - \mu = 0
\end{aligned}
$$

- The MLE of the mean of a Gaussian is UNBIASED

Maximum-Likelihood Estimation (MLE) of a Gaussian

- Are *all* MLEs unbiased? No!

$$\hat{\sigma}^2(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

$$\mathbb{E}_{\mathbf{X}}\left[\hat{\sigma}^2(\mathbf{X}) - \sigma^2\right] = \ldots = \frac{N-1}{N}\sigma^2 - \sigma^2 = -\frac{1}{N}\sigma^2$$

- The MLE of the variance of a Gaussian is BIASED
- We can easily get an unbiased estimator

$$\tilde{\sigma}^2(\mathbf{X}) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Bias and Variance (Regression Example)

- Estimator $\hat{f}_{\mathcal{D}}$ from training data $\mathcal{D}$, data generated by

$$y(\mathbf{x}_q) = f(\mathbf{x}_q) + \epsilon$$

with $E\{\epsilon\} = 0$ and $\mathrm{Var}\{\epsilon\} = \sigma_\epsilon^2$

- Note $f(\mathbf{x})$ is not random

$$E_{\mathcal{D},\epsilon}\{y(\mathbf{x}_q)\} = E_{\mathcal{D},\epsilon}\{f(\mathbf{x}_q)\} = f(\mathbf{x}_q)$$

- **Expected Squared Error** for query $\mathbf{x}_q$ estimated from all possible data sets $\mathcal{D}$

$$L_{\hat{f}}(\mathbf{x}_q) = E_{\mathcal{D},\epsilon}\left\{\left(y(\mathbf{x}_q) - \hat{f}_{\mathcal{D}}(\mathbf{x}_q)\right)^2\right\}$$

$$= E_{\mathcal{D},\epsilon}\left\{\left(y(\mathbf{x}_q) - f(\mathbf{x}_q) + f(\mathbf{x}_q) - \hat{f}_{\mathcal{D}}(\mathbf{x}_q)\right)^2\right\}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Bias and Variance (Regression Example)

Expected Squared Error for query $\mathbf{x}_q$ estimated from all possible data sets $\mathcal{D}$

$$L_{\hat{f}}(\mathbf{x}_q) = E_{\mathcal{D},\epsilon}\left\{ \underbrace{(y(\mathbf{x}_q) - f(\mathbf{x}_q))^2}_{=\epsilon^2} + \left(f(\mathbf{x}_q) - \hat{f}_{\mathcal{D}}(\mathbf{x}_q)\right)^2 + 2\underbrace{(y(\mathbf{x}_q) - f(\mathbf{x}_q))(f(\mathbf{x}_q) - \hat{f}_{\mathcal{D}}(\mathbf{x}_q))}_{=\epsilon\left(f(\mathbf{x}_q) - \hat{f}_{\mathcal{D}}(\mathbf{x})\right)} \right\}$$

$$= \sigma_\epsilon^2 + E_{\mathcal{D}}\left\{ \left(f(\mathbf{x}_q) - \hat{f}_{\mathcal{D}}(\mathbf{x}_q)\right)^2 \right\} = \sigma_\epsilon^2 + \mathrm{bias}^2\left\{\hat{f}_{\mathcal{D}}(\mathbf{x}_q)\right\} + \mathrm{var}\left\{\hat{f}_{\mathcal{D}}(\mathbf{x}_q)\right\}$$

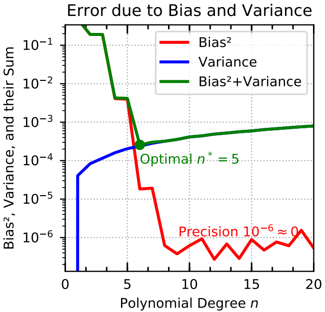
using $\bar{\hat{f}}(\mathbf{x}_q) = E_{\mathcal{D}}\left\{\hat{f}_{\mathcal{D}}(\mathbf{x}_q)\right\}$, we obtain

$$E_{\mathcal{D}}\left\{ \left(f(\mathbf{x}_q) - \hat{f}_{\mathcal{D}}(\mathbf{x}_q)\right)^2 \right\} = E_{\mathcal{D}}\left\{ \left(f(\mathbf{x}_q) - \bar{\hat{f}}(\mathbf{x}_q) + \bar{\hat{f}}(\mathbf{x}_q) - \hat{f}_{\mathcal{D}}(\mathbf{x}_q)\right)^2 \right\}$$

$$= \underbrace{\left(f(\mathbf{x}_q) - \bar{\hat{f}}(\mathbf{x}_q)\right)^2}_{=\mathrm{bias}^2\left\{\hat{f}_{\mathcal{D}}(\mathbf{x}_q)\right\}} + \underbrace{E_{\mathcal{D}}\left\{ \left(\bar{\hat{f}}(\mathbf{x}_q) - \hat{f}_{\mathcal{D}}(\mathbf{x}_q)\right)^2 \right\}}_{=\mathrm{var}\left\{\hat{f}_{\mathcal{D}}(\mathbf{x}_q)\right\}}$$

Bias-Variance Tradeoff

- (Total) **Bias** $\text{bias}^2\left\{\hat{f}_{\mathcal{D}}\right\} = E_{\mathbf{x}_q}\left\{\left(f(\mathbf{x}_q) - E_{\mathcal{D}}\{\hat{f}_{\mathcal{D}}(\mathbf{x}_q)\}\right)^2\right\}$
  - Structure error
  - Model $\hat{f}_{\mathcal{D}}(\mathbf{x}_q)$ cannot do better
- (Total) **Variance** $\text{var}\left\{\hat{f}_{\mathcal{D}}(\mathbf{x}_q)\right\} = E_{\mathbf{x}_q,\mathcal{D}}\left\{\left(\hat{f}_{\mathcal{D}}(\mathbf{x}_q) - E_{\tilde{\mathcal{D}}}\{\hat{f}_{\tilde{\mathcal{D}}}(\mathbf{x}_q)\}\right)^2\right\}$
  - Estimation error
  - Finite data sets will always have errors
- **Expected Total Error** $\propto$ Bias^2+Variance
  - You typically cannot minimize both



Bias-Variance Tradeoff

- Our learning algorithm will only generalize well if we find the right tradeoff between bias and variance
  - Simple enough to prevent *overfitting* to the particular training data set that we have
  - Yet expressive enough to be able to represent the important properties of the data
- To ensure that our learning algorithm works well, we have to evaluate it on test data
  - But what if we don't have any?

TECHNISCHE UNIVERSITÄT DARMSTADT

How do choose the model?

- **Goal**: Find a good model (e.g., good set of features)
- **Split the dataset into:**

1. *Training Set*: Fit Parameters
2. *Validation Set*: Choose model class or single parameters
3. *Test Set*: Estimate prediction error of trained model

$\Rightarrow$ **Error/loss L needs to be estimated on an independent set!**

Model Selection: Cross Validation

- Partition data into K sets $\mathcal{D}_\kappa$, use $K-1$ for *Training* and 1 for *Validation*

and compute

$$\theta_k(\mathcal{M}_j) = \mathbf{argmin}_{\theta \in \mathcal{M}_j} \sum_{\kappa \neq k} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_\kappa} L_{f_{\boldsymbol{\theta}}}(\mathbf{x}_i, y_i)$$

$$L_k(\mathcal{M}_j) = \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_k} L_{f_{\boldsymbol{\theta}k}}(\mathbf{x}_i, y_i)$$

- **Exhaustive Cross Validation**: Try all partitioning possibilities $\Rightarrow$ **Computationally expensive**

- **Bootstrap**: Randomly sample non-overlapping training / validation sets

TECHNISCHE UNIVERSITÄT DARMSTADT

Model Selection: K-fold Cross Validation

- **Cheapest reasonable approach: K-fold Cross Validation**

Training Set	Training Set	Validation Set
Training Set	Validation Set	Training Set
Validation Set	Training Set	Training Set

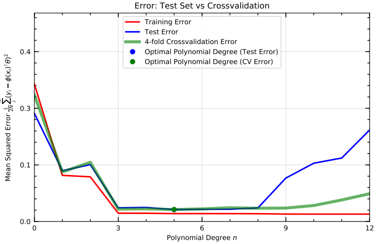
- Compute the validation loss and choose Model

$$\mathcal{M}^* = \operatorname{argmin}_{\mathcal{M}} \frac{1}{K} \sum_{k=1}^{K} L_k(\mathcal{M})$$

  with smallest average validation loss
- **Leave-one-out cross-validation (LOOCV)**: $K = N - 1 \Rightarrow$ Validation set size 1

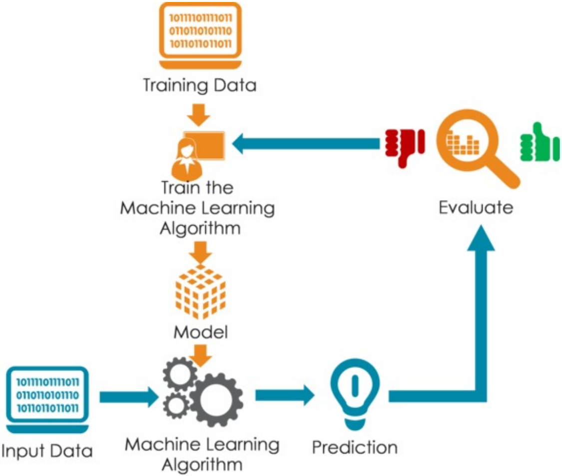
Model Selection: K-fold Cross Validation

4-fold Cross-Validation



Leave-one-out Cross-Validation

Machine Learning Cycle



[https://devpost.com]

TECHNISCHE UNIVERSITÄT DARMSTADT

Outline

3. Wrap-Up

You know now:

- Different algorithms for clustering
- How to compute the bias and variance of an estimator
- What the Bias-Variance tradeoff is
- What MVUE and BLUE mean
- The difference between unbiased and biased estimators
- How to mimic test data evaluation using cross-validation

Self-Test Questions

- How can we find meaningful clusters in the data?
- How does density estimation with mixture models relate to clustering?
- What is the bias-variance trade-off?
- What is a BLUE estimator?
- Are maximum likelihood estimators always unbiased?
- What is leave one out cross-validation? What do we need it for?

Homework

- Reading Assignment for next lecture
  - Murphy ch. 7
  - Bishop ch. 3