Elements of Machine Learning & Data Science

Winter semester 2023/24

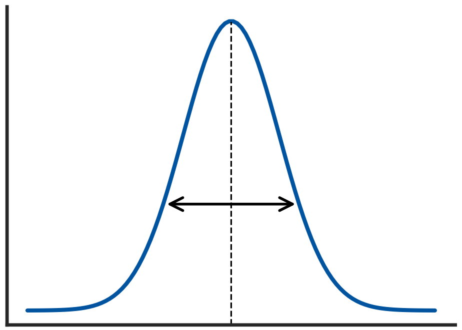
Lecture 4 – Probability Density Estimation I

20.10.2023

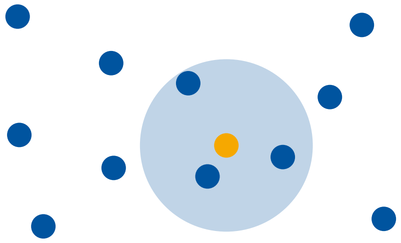
Prof. Bastian Leibe

Machine Learning Topics

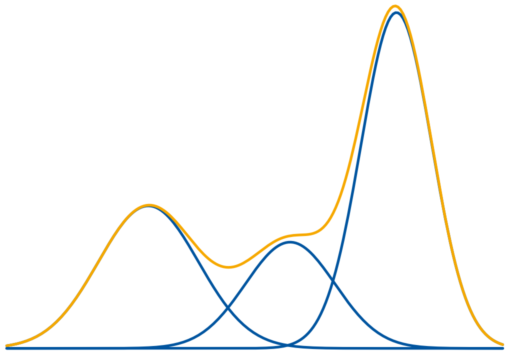
1. Introduction to ML
2. **Probability Density Estimation**
3. Linear Discriminants
4. Linear Regression
5. Logistic Regression
6. Support Vector Machines
7. AdaBoost
8. Neural Network Basics



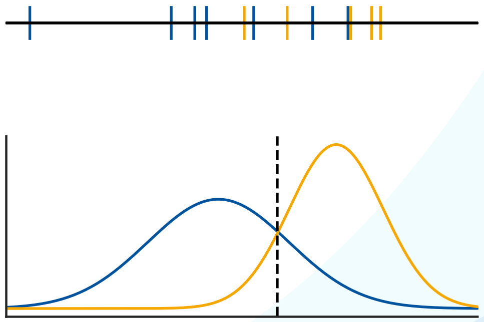
Parametric Methods
& ML-Algorithm



Nonparametric Methods



Mixtures of Gaussians
& EM-Algorithm



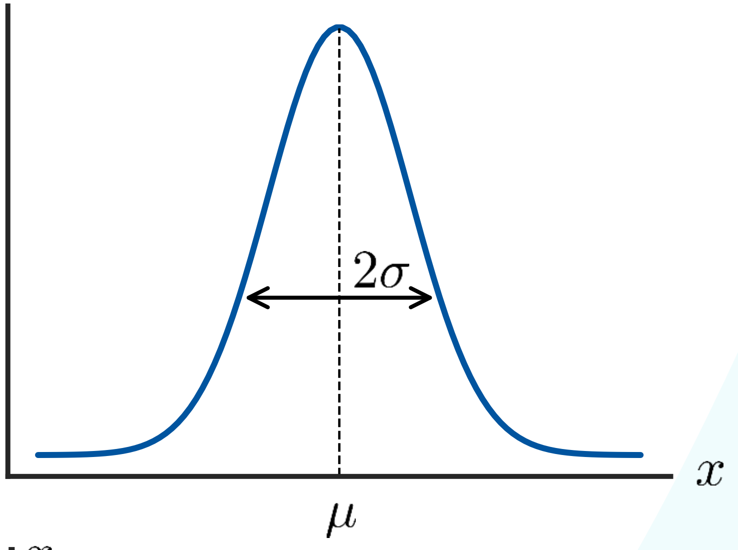
Bayes Classifiers

Recap: The Gaussian (or Normal) Distribution

- One-dimensional (univariate) case:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

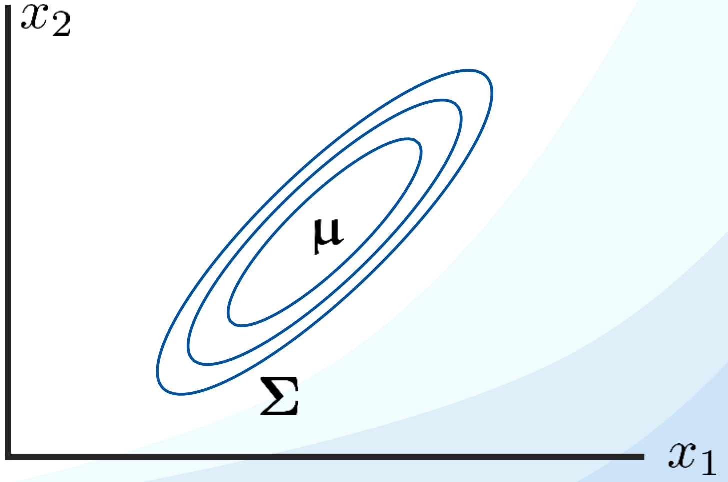
Mean   Variance



- Multi-dimensional (multivariate) case:

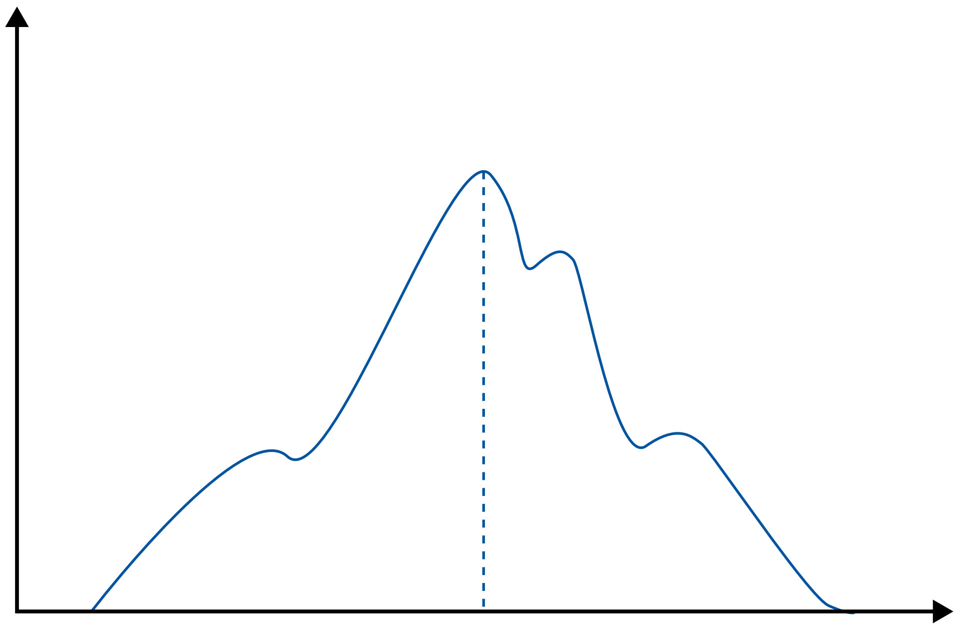
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Mean vector   Covariance matrix



Probability Density Estimation

1. Probability Distributions
2. **Parametric Methods**
3. Nonparametric Methods
4. Mixture Models
5. Bayes Classifier
6. K-NN Classifier

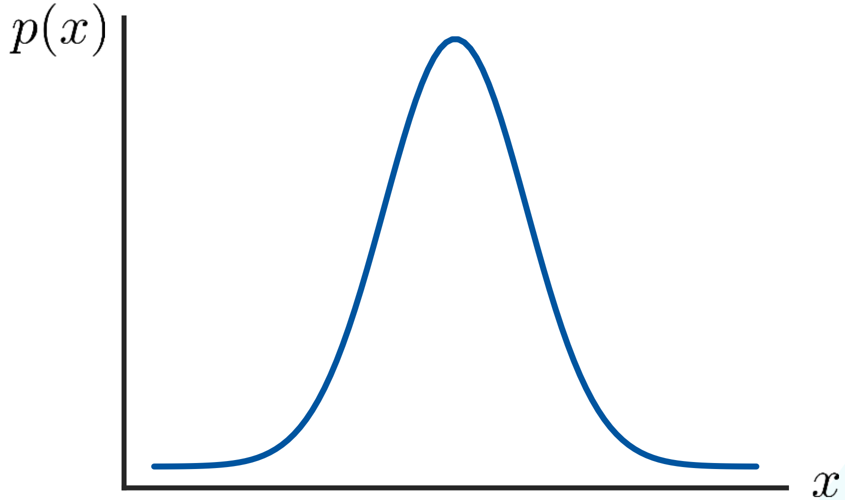


Parametric Methods

- In parametric methods, we assume that we know the parametric form of the underlying data distribution.
  - I.e., the equation of the pdf with parameters θ.

Example: $p(x) = \mathcal{N}(x|\mu, \sigma)$

$$\theta = (\mu, \sigma)$$



Parametric Methods

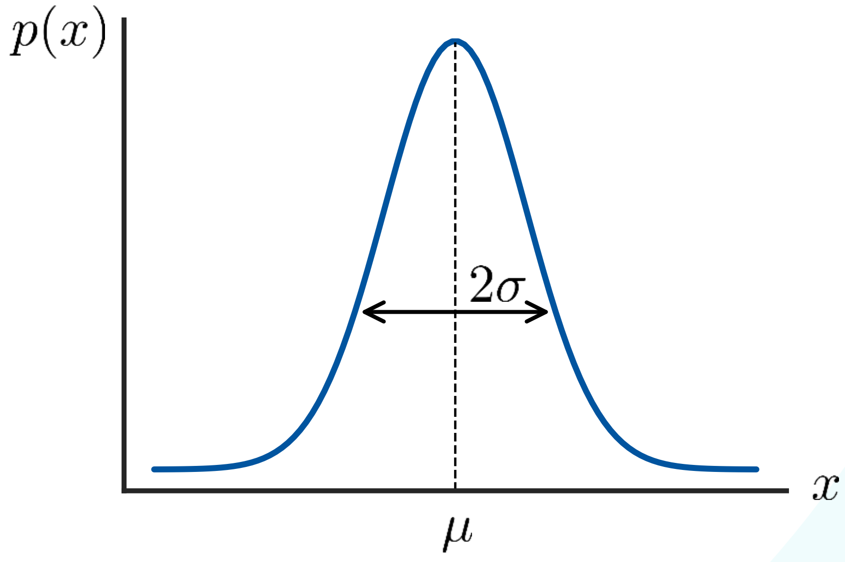
- In parametric methods, we assume that we know the parametric form of the underlying data distribution.
  - I.e., the equation of the pdf with parameters θ.
- Goal: Estimate θ from training data $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$.
- Likelihood of θ:

$$L(\theta) = p(\mathcal{X}|\theta)$$

Probability that the data $\mathcal{X}$ was indeed generated by a distribution with parameters θ.

Example: $p(x) = \mathcal{N}(x|\mu, \sigma)$

$\theta = (\mu, \sigma)$



Maximum Likelihood Approach

- Idea: Find optimal parameters by maximizing $L(\theta)$.
- Computation of the likelihood:
  - Single data point (e.g., for Gaussian):

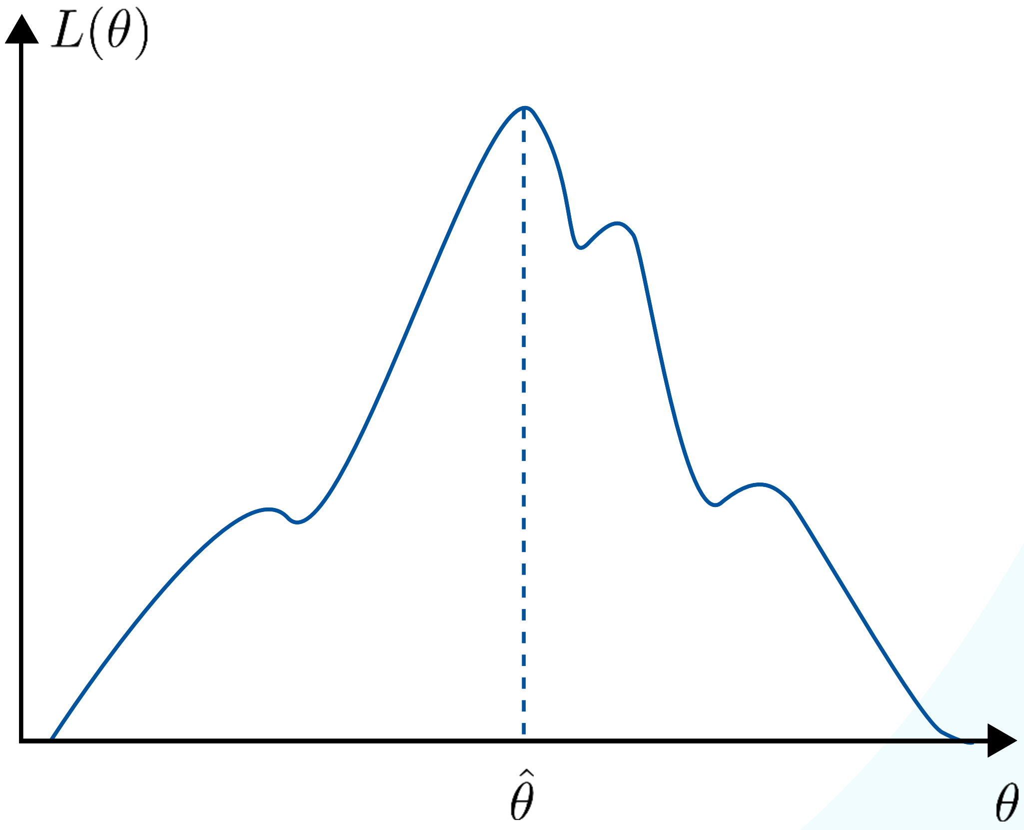
$$p(x_n|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

  - Assumption: all data points are independent

$$L(\theta) = p(\mathcal{X}|\theta) = \prod_{n=1}^{N} p(x_n|\theta)$$

  - Negative Log-Likelihood ("Energy"):

$$E(\theta) = -\ln L(\theta) = -\sum_{n=1}^{N} \ln p(x_n|\theta)$$



Maximizing the likelihood ⇔ minimizing the negative log-likelihood.

- Minimizing the negative log-likelihood:
  - Take the derivative and set it to zero.

$$\frac{\partial}{\partial\theta}E(\theta) = -\frac{\partial}{\partial\theta}\sum_{n=1}^{N}\ln p(x_n|\theta) = -\sum_{n=1}^{N}\frac{\frac{\partial}{\partial\theta}p(x_n|\theta)}{p(x_n|\theta)} \stackrel{!}{=} 0$$

  - Log-likelihood for Normal distribution (1D case):

$$p(x_n|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

$$\frac{\partial}{\partial\mu}p(x_n|\mu,\sigma) = -\frac{2(x_n-\mu)}{2\sigma^2}p(x_n|\mu,\sigma)$$

$$\frac{\partial}{\partial\mu}E(\mu,\sigma) = -\sum_{n=1}^{N}\frac{\frac{\partial}{\partial\mu}p(x_n|\mu,\sigma)}{p(x_n|\mu,\sigma)}$$

$$= -\sum_{n=1}^{N}-\frac{2(x_n-\mu)}{2\sigma^2}\frac{p(x_n|\mu,\sigma)}{p(x_n|\mu,\sigma)}$$

$$= \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) = \frac{1}{\sigma^2}\left(\sum_{n=1}^{N}x_n - N\mu\right) \stackrel{!}{=} 0 \quad \Longleftrightarrow \quad \hat{\mu} = \frac{1}{N}\sum_{n=1}^{N}x_n$$

- By minimizing the negative log-likelihood, we found:

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

sample mean

Similarly, we can derive:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N} (x_n - \hat{\mu})^2$$

sample variance

- $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$ is the Maximum Likelihood estimate for the parameters of a Gaussian distribution.
  - This is a very important result.
  - Unfortunately, it is wrong…

- To be precise, the result is not wrong, but biased.
- Assume the samples $x_1, x_2, \ldots, x_N$ come from a true Gaussian distribution with mean μ and variance σ^2
  - It can be shown that the expected estimates are then

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2$$

$\Rightarrow$ *The ML estimate will underestimate the true variance!*

- We can correct for this bias:

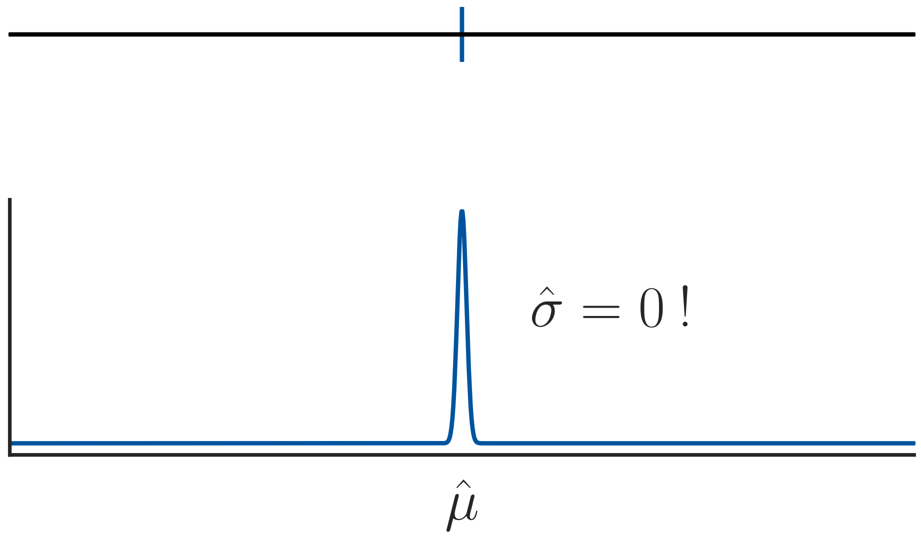
$$\hat{\sigma}^2 = \frac{N}{N-1}\sigma^2_{\mathrm{ML}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$

- Maximum Likelihood has several significant limitations.
  - It systematically underestimates the variance of the distribution!
  - E.g., consider the estimate for a single sample:

$$N = 1,\ \mathcal{X} = \{x_1\}$$

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n = x_1$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2 = 0$$

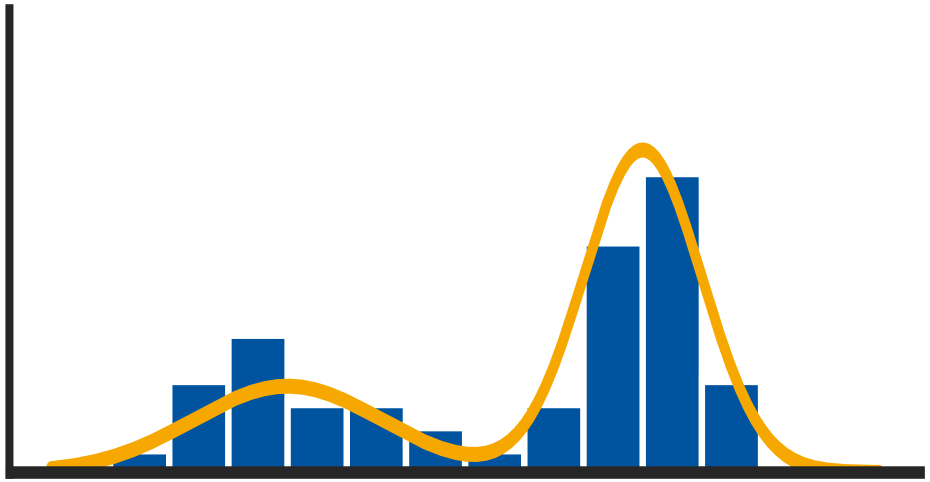


- We say ML overfits to the observed data.

- *We will still often use Maximum Likelihood, but it is important to know about this effect.*

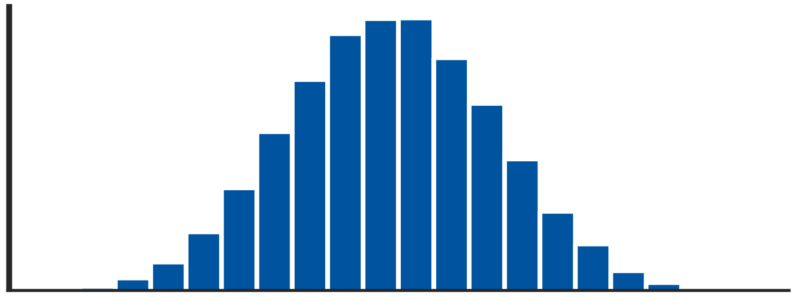
Probability Density Estimation

1. Probability Distributions
2. Parametric Methods
3. **Nonparametric Methods**
   a) **Histograms**
   b) Kernel Methods & k-Nearest Neighbors
4. Mixture Models
5. Bayes Classifier
6. K-NN Classifier

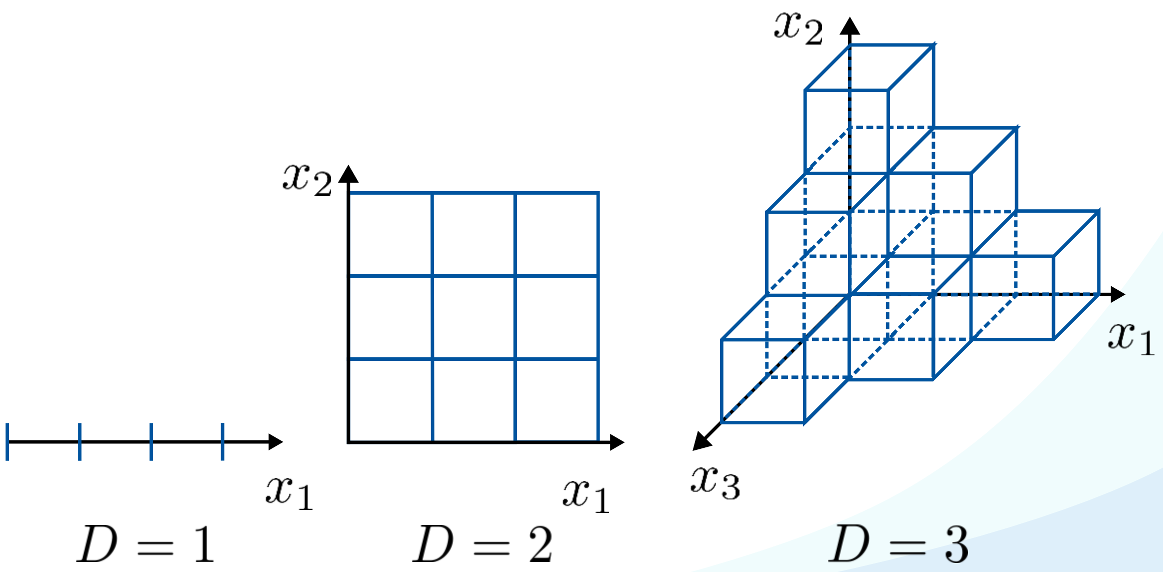


Histograms

- Partition the data space into M distinct bins with widths Δ_i and count the number of observations n_i in each bin.
- Then, $p_i = \dfrac{n_i}{N\Delta_i}$.
- Often the same width is used for all bins.
- This can be done, in principle, for any dimensionality D.

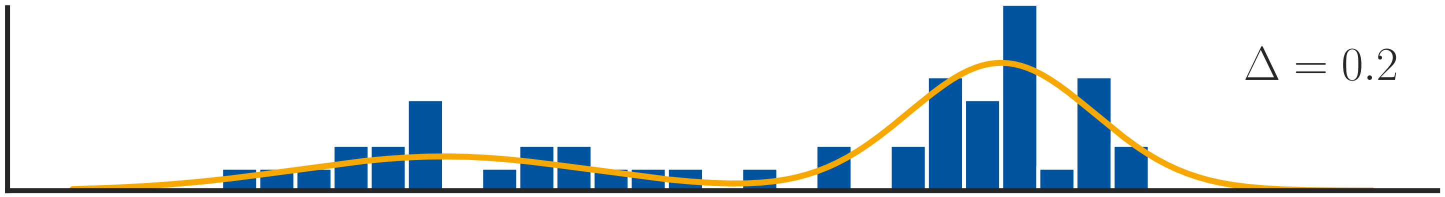


…but the required number of bins grows exponentially with D!



The bin width Δ acts as a smoothing factor.

Not smooth enough



About ok

Too smooth

Advantages

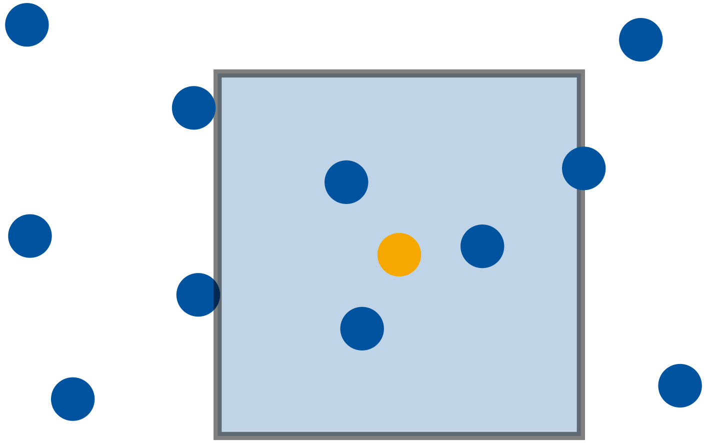
- Very general method. In the limit $(N \to \infty)$, every probability density can be represented.
- No need to store the data points once histogram is computed.

Limitations

- Rather brute-force.
- Discontinuities at bin edges.
- Choosing right bin size is hard.
- Unsuitable for high-dimensional feature spaces.

Probability Density Estimation

1. Probability Distributions
2. Parametric Methods
3. **Nonparametric Methods**
   a) Histograms
   b) **Kernel Methods & k-Nearest Neighbors**
4. Mixture Models
5. Bayes Classifier
6. K-NN Classifier



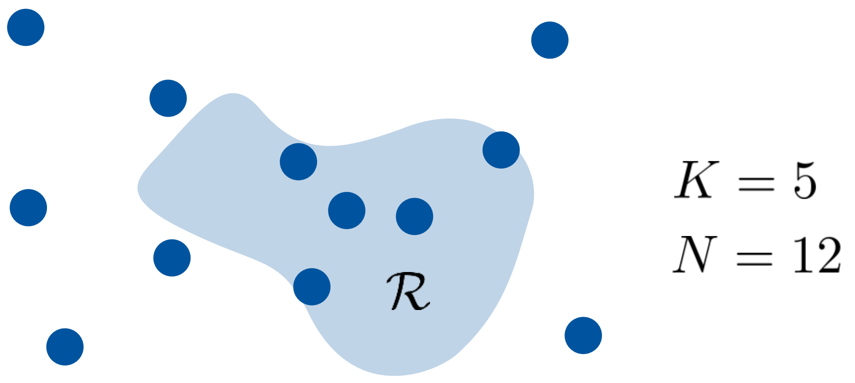
Kernel Methods and k-Nearest Neighbors

- Data point $\mathbf{x}$ comes from pdf $p(\mathbf{x})$.
  - Probability that $\mathbf{x}$ falls into small region $\mathcal{R}$:

$$P = \int_{\mathcal{R}} p(y)dy \approx p(\mathbf{x})V$$

- Estimate $p(\mathbf{x})$ from samples
  - Let K be the number of samples that fall into $\mathcal{R}$.
  - If the number of samples N is sufficiently large, we can estimate P as:

$$P = \frac{K}{N} \quad \Rightarrow \quad p(\mathbf{x}) \approx \frac{K}{NV}$$



For sufficiently small $\mathcal{R}$, $p(\mathbf{x})$ is roughly constant.

V : volume of $\mathcal{R}$.

$$p(\mathbf{x}) \approx \frac{K}{NV}$$

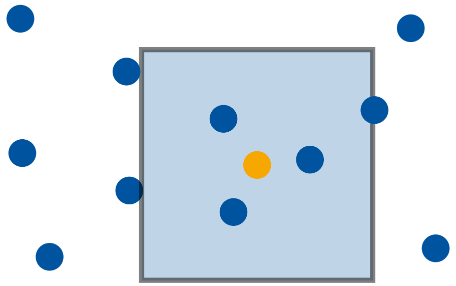
fixed V
determine K

fixed K
determine V

Kernel Methods

k-Nearest Neighbors

Example: Determine the number K of data points inside a fixed hypercube



Kernel Methods

- Hypercube of dimension D with edge length h:

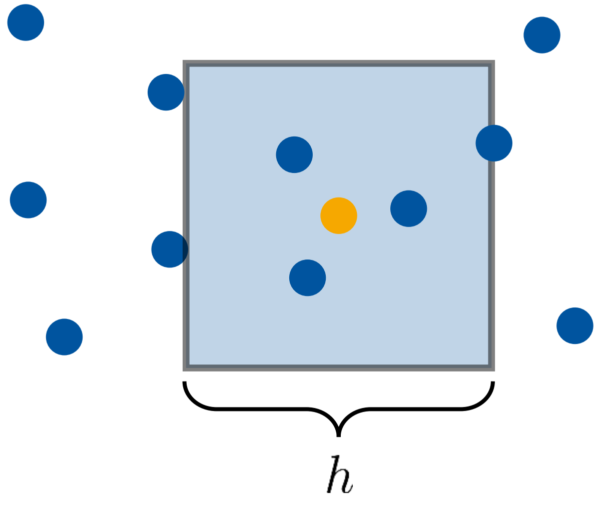
$$k(\mathbf{u}) = \begin{cases} 1, & \text{if } |u_i| \leq \frac{1}{2}h, \; i = 1, \ldots, D \\ 0, & \text{otherwise} \end{cases}$$

$$K = \sum_{n=1}^{N} k(\mathbf{x} - \mathbf{x}_n) \qquad V = \int k(\mathbf{u}) d\mathbf{u} = h^D$$

- Probability density estimate:

$$p(\mathbf{x}) \approx \frac{K}{NV} = \frac{1}{Nh^D} \sum_{n=1}^{N} k(\mathbf{x} - \mathbf{x}_n)$$

- This method is known as Parzen Window estimation.



- In general, we can use any kernel such that

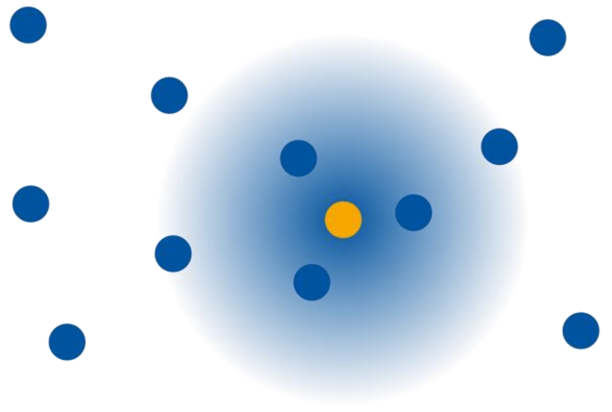
$$k(\mathbf{u}) \geq 0, \quad \int k(\mathbf{u}) d\mathbf{u} = 1$$

$$K = \sum_{n=1}^{N} k(\mathbf{x} - \mathbf{x}_n)$$

- Then, we get the probability density estimate

$$p(\mathbf{x}) \approx \frac{K}{NV} = \frac{1}{N} \sum_{n=1}^{N} k(\mathbf{x} - \mathbf{x}_n)$$

- This is known as Kernel Density Estimation.



E.g., a Gaussian kernel for smoother boundaries.

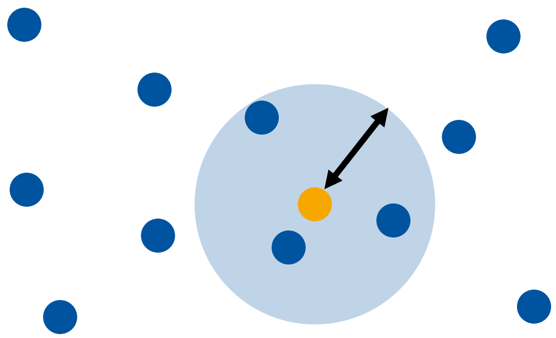
$$p(\mathbf{x}) \approx \frac{K}{NV}$$

fixed V
determine K

fixed K
determine V

Kernel Methods

k-Nearest Neighbors



Increase the volume V until the K next data points are found.

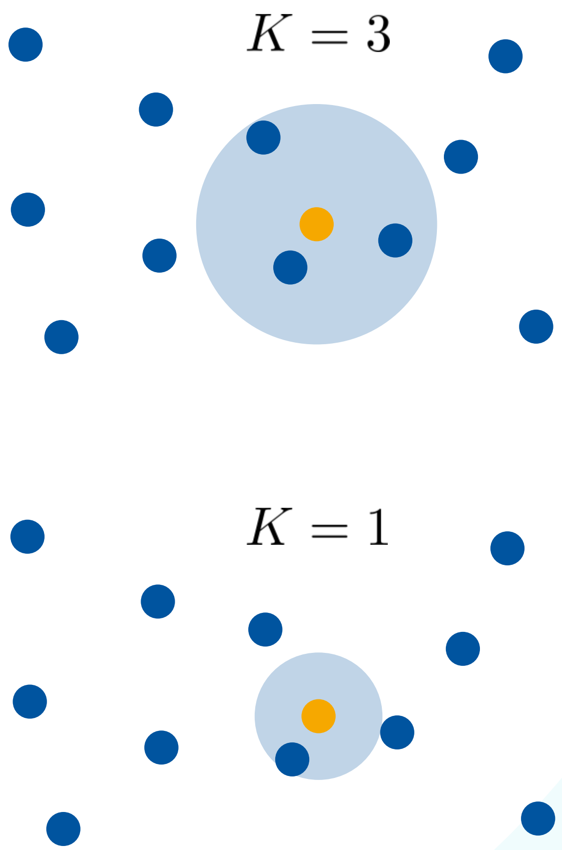
k-Nearest Neighbors

- Fix K, estimate V from the data.
- Consider a hypersphere centered on $\mathbf{x}$ and let it grow to a volume V^* that includes K of the given N data points.
- Then

$$p(\mathbf{x}) \approx \frac{K}{NV^*}$$

- Side note:
  - Strictly speaking, the model produced by k-NN is not a true density model, because the integral over all space diverges.
  - E.g. consider $K = 1$ and a sample exactly on a data point.

$$V^* = 0 \qquad \Rightarrow \qquad p(\mathbf{x}) \approx \frac{K}{N \cdot 0}$$



Advantages

- Very general. In the limit $(N \to \infty)$, every probability density can be represented.
- No computation during training phase
  - Just need to store training set

Limitations

- Requires storing and computing with the entire dataset.
  - Computational costs linear in the number of data points.
  - Can be improved through efficient storage structures (at the cost of some computation during training).
- Choosing the kernel size/K is a hyperparameter optimization problem.

Bias-Variance Tradeoff

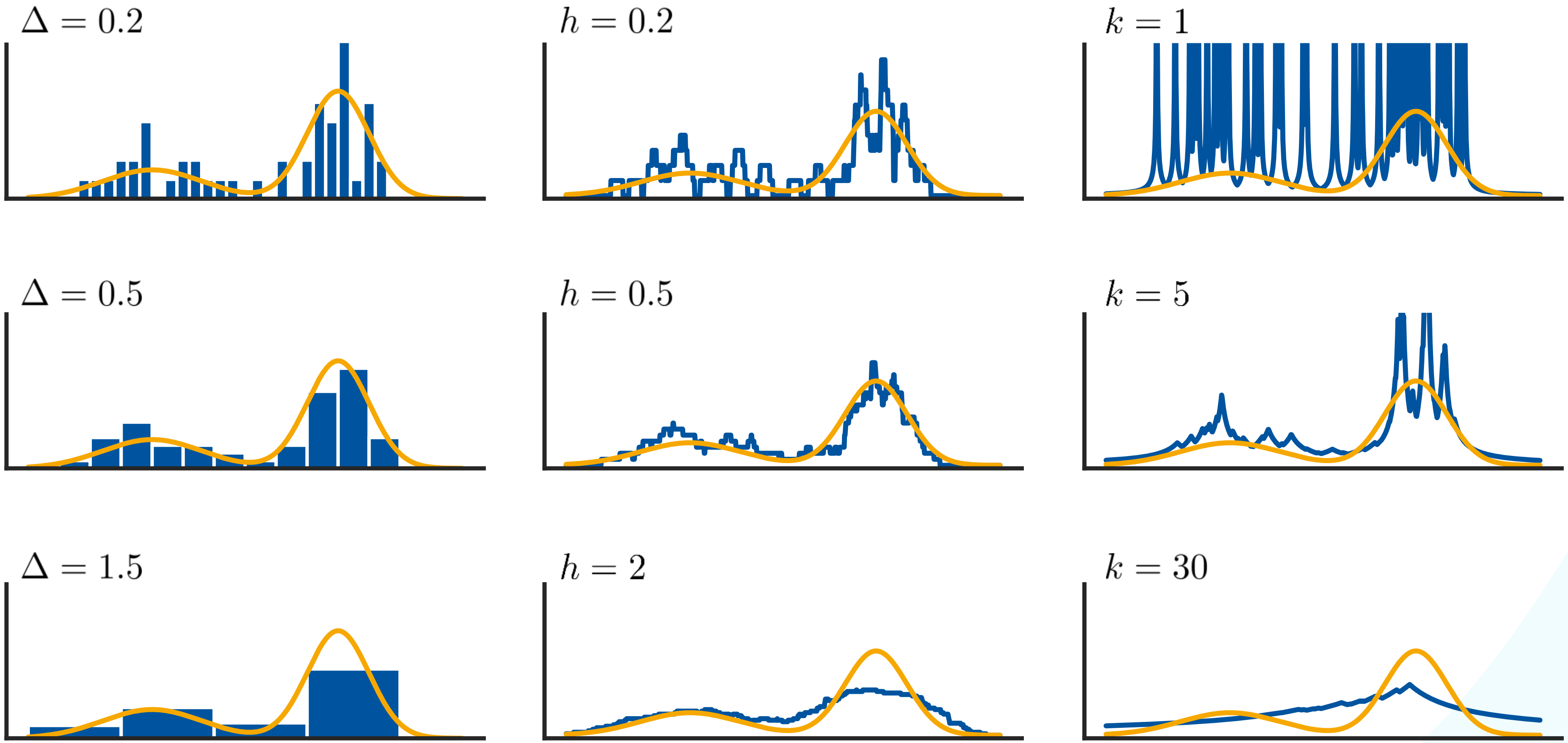
Not smooth enough

Too much variance

About ok

Too smooth

Too much bias



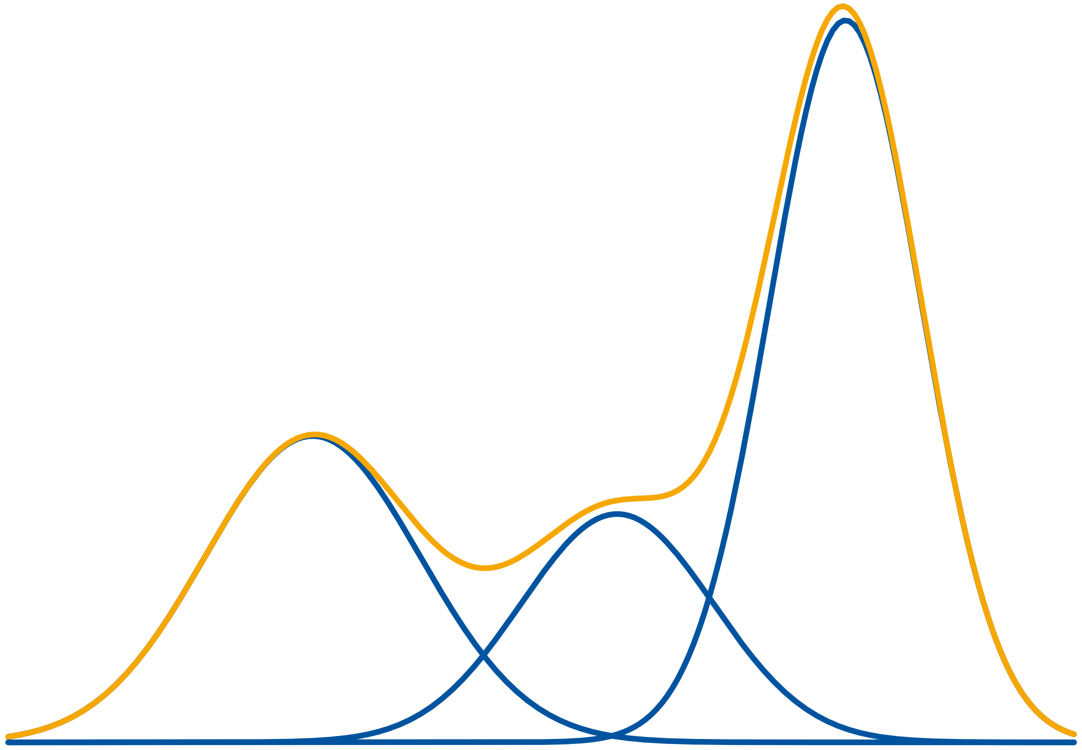
Histograms: Bin width Δ

Parzen Window: Kernel size h

k-NN: # of neighbors k

Probability Density Estimation

1. Probability Distributions
2. Parametric Methods
3. Nonparametric Methods
4. **Mixture Models**
5. Bayes Classifier
6. K-NN Classifier



References and Further Reading

- More information in Bishop's book
  - Gaussian distribution and ML: Ch. 1.2.4 and 2.3.1-2.3.4.
  - Nonparametric methods: Ch. 2.5.

Christopher M. Bishop
Pattern Recognition and Machine Learning
Springer, 2006