Machine Intelligence 2

5.1 Probability Density Estimation

Prof. Dr. Klaus Obermayer

Fachgebiet Neuronale Informationsverarbeitung (NI)

SS 2018

Density estimation

Density estimation is relevant

If $p(x)$ is known, all predictable quantities can be deduced (mean, variance, higher order moments, $p(x$ in interval$)$...

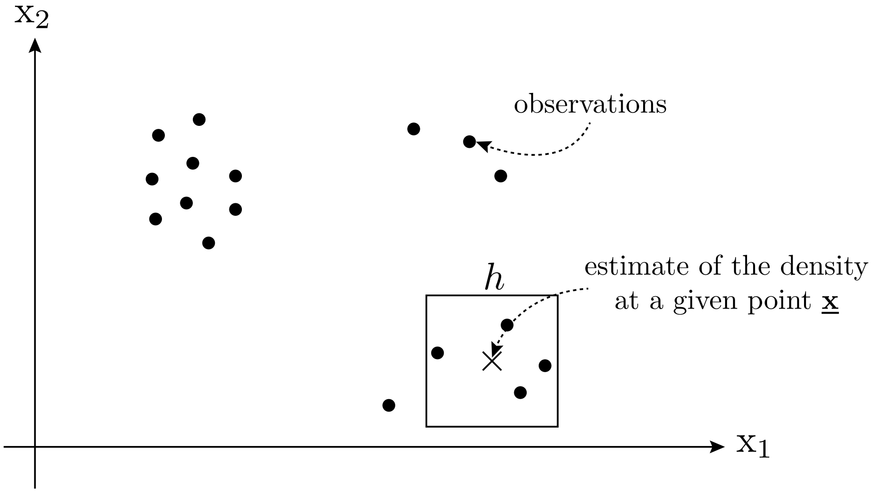
Density estimation is difficult (without prior knowledge)

How can we estimate the density for each possible outcome?

2 strategies:

1. **parametric** methods: model-based (e.g. Gaussian densities)
2. **nonparametric** methods: data driven (cf. Kernel density estimate)

(Nonparametric) Kernel density estimation



"Gliding histograms"

Count the number of data points within a volume V centered on $\underline{\mathbf{x}}$.

Histogram kernel:

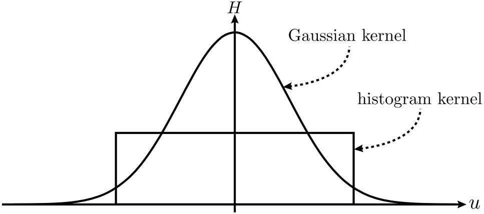
$$H(\underline{\mathbf{u}}) = \begin{cases} 1, & |u_j| < \frac{1}{2}, \forall j \in 1, \dots, n \\ 0, & \text{else} \end{cases}$$

Density estimate ("gliding histogram"):

$$\widehat{P}(\underline{\mathbf{x}}) = \underbrace{\frac{1}{h^n}}_{\substack{\text{normalization} \\ \text{("density"!)}}} \cdot \underbrace{\frac{1}{p} \overbrace{\sum_{\alpha=1}^{p} H\left(\frac{\underline{\mathbf{x}} - \underline{\mathbf{x}}^{(\alpha)}}{h}\right)}^{\substack{\text{number of data points} \\ \text{within volume } V \text{ around } \underline{\mathbf{x}}}}}_{\text{fraction of data points}}$$

Histogram kernels lead to discontinuous pdf estimates $\rightsquigarrow$ use other kernels for smooth pdf estimates.

Gaussian kernels



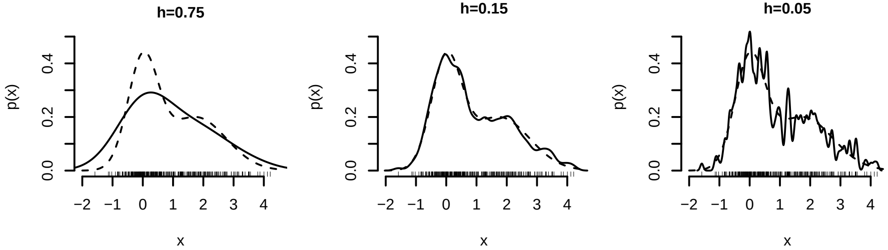
Gaussian kernel:

$$H(\underline{\mathbf{u}}) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(- \frac{\underline{\mathbf{u}}^2}{2} \right)$$

Density estimate:

$$\begin{aligned}
\widehat{P}(\underline{\mathbf{x}}) &= \frac{1}{h^n} \cdot \frac{1}{p} \sum_{\alpha=1}^{p} H\left(\frac{\underline{\mathbf{x}} - \underline{\mathbf{x}}^{(\alpha)}}{h} \right) \\
&= \frac{1}{p} \sum_{\alpha=1}^{p} \frac{1}{\left(2\pi h^2\right)^{\frac{n}{2}}} \exp\left\{ - \frac{\left(\underline{\mathbf{x}} - \underline{\mathbf{x}}^{(\alpha)}\right)^2}{2h^2} \right\}
\end{aligned}$$

Effects of kernel width



Choice of kernel width $\Rightarrow$ model selection / validation

Parametric density estimation

observations: $\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \alpha = 1, \dots, p$
parametrized family of pdfs: $\widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}}) \leftarrow$ "generative model"

example: multivariate Gaussian

$$\widehat{P}(\underline{\mathbf{x}}; \overbrace{\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\Sigma}}}^{\underline{\mathbf{w}}}) = \underbrace{\frac{1}{\sqrt{(2\pi)^N \det \underline{\boldsymbol{\Sigma}}}} \exp\left(-\frac{1}{2}(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}})^T \underline{\boldsymbol{\Sigma}}^{-1}(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}})\right)}_{\mathcal{N}(\underline{\mathbf{x}}; \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\Sigma}})}$$

comment

here: $\widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}})$ for unconditional densities $P(\underline{\mathbf{x}}) \Rightarrow$ unsupervised learning
MI I: $\widehat{P}(y|\underline{\mathbf{x}}; \underline{\mathbf{w}})$ for conditional densities $P(y|\underline{\mathbf{x}}) \Rightarrow$ supervised learning

$\Rightarrow$ model selection

Parametric density estimation

Generative model: parametrized family of pdfs: $\widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}})$

Model selection

Select the model (set of parameters) which is most similar to the true density!

Kullback-Leibler-Divergence

$$\mathrm{D_{KL}}\Big[P(\underline{\mathbf{x}}), \widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}})\Big] = \int d\underline{\mathbf{x}} P(\underline{\mathbf{x}}) \ln \frac{P(\underline{\mathbf{x}})}{\widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}})} \stackrel{!}{=} \min_{(\underline{\mathbf{w}})}$$

- $\mathrm{D_{KL}} \geq 0$ and $\mathrm{D_{KL}} = 0$ iff $\widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}}) = P(\underline{\mathbf{x}})$
- distance measure between probability distributions

Model selection via Empirical Risk Minimization

$$\mathrm{D}_{\mathrm{KL}}(P, \hat{P}_{\underline{\mathbf{w}}}) \stackrel{!}{=} \min_{(\underline{\mathbf{w}})}$$

$$\begin{aligned} \underline{\mathbf{w}}^* &= \underset{(\underline{\mathbf{w}})}{\operatorname{argmin}} \left\{ \int d\underline{\mathbf{x}} P(\underline{\mathbf{x}}) \ln P(\underline{\mathbf{x}}) - \int d\underline{\mathbf{x}} P(\underline{\mathbf{x}}) \ln \widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}}) \right\} \\ &= \underset{(\underline{\mathbf{w}})}{\operatorname{argmin}} \left\{ \underbrace{- \int d\underline{\mathbf{x}} P(\underline{\mathbf{x}}) \ln \widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}})}_{E^G_{[\underline{\mathbf{w}}]}} \right\} \end{aligned}$$

"cross entropy"

$$E^G \stackrel{!}{=} \min_{(\underline{\mathbf{w}})}$$

Problem: $P(\underline{\mathbf{x}})$ is unknown.

Model selection via Empirical Risk Minimization

mathematical expectation E^G	$\longrightarrow$	empirical average E^T
"generalization cost"		"training cost"

cost function:

$$E^T = -\frac{1}{p} \sum_{\alpha=1}^{p} \ln \widehat{P}\big(\underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}\big)$$

When is this a reasonable procedure? $\leadsto$ statistical learning theory *(MI I)*

criterion for model selection

$$\boxed{E^T = -\tfrac{1}{p} \sum_{\alpha=1}^{p} \ln \widehat{P}\big(\underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}\big) \stackrel{!}{=} \min_{(\underline{\mathbf{w}})}}$$

Optimization of the empirical risk

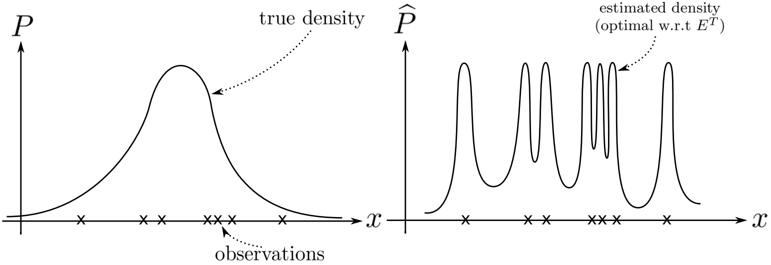
$$\underbrace{E^T_{[\underline{\mathbf{w}}]}}_{\text{total cost}} = -\frac{1}{p}\sum_{\alpha=1}^{p} \ln \widehat{P}\big(\underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}\big) = \frac{1}{p}\sum_{\alpha=1}^{p} \underbrace{e^{(\alpha)}_{[\underline{\mathbf{w}}]}}_{\text{individual cost}}$$

standard procedures e.g. (stochastic) gradient descent – cf. MI I

$$\left.\begin{array}{ll} \text{"batch"-learning:} & \Delta\underline{\mathbf{w}} = -\varepsilon \frac{\partial E^T}{\partial \underline{\mathbf{w}}} \\[2ex] \text{"on-line"-learning:} & \Delta\underline{\mathbf{w}} = -\varepsilon \frac{\partial e^{(\alpha)}}{\partial \underline{\mathbf{w}}} \end{array}\right\} \text{examples for gradient-based methods}$$

Validation

Minimized training cost underestimates the corresponding generalization cost



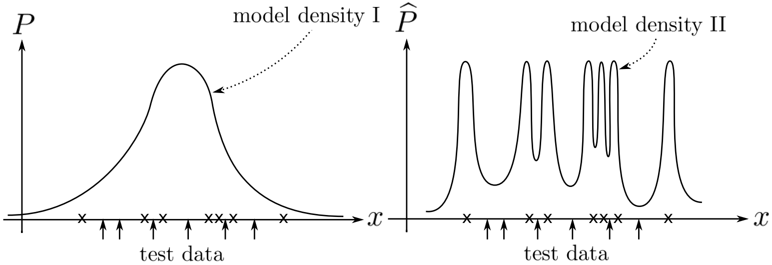
Overfitting:
E^T small but E^G large $\Rightarrow$ test-set method, n-fold cross-validation

Test-set method

$$\text{observations: } \begin{cases} \text{training data} & \{\underline{\mathbf{x}}^{(\alpha)}\}, \alpha = 1, \ldots, p \\ \\ \text{test data} & \{\underline{\mathbf{x}}^{(\beta)}\}, \beta = 1, \ldots, q \end{cases}$$

$$\widehat{E}^G = \frac{1}{q} \sum_{\beta=1}^{q} e^{(\beta)} \leftarrow \text{ estimate of } E^G$$

Test-set method



- $E^T_{(I)} > E^T_{(II)}$ but $E^G_{(I)} << E^G_{(II)}$
- Alternative method: **n-fold cross-validation** *(MI I)*

Comment

Validation methods are essential for estimating hyperparameters for non-parametric methods (e.g. Kernel density estimate).

The likelihood function

generative model

$\widehat{P}(\underline{\mathbf{x}}; \underline{\mathbf{w}})$ probability density for the generation of one data point

likelihood of the model = p(observations given the model)

assuming iid. observations:

$$\widehat{P}\Big(\big\{\underline{\mathbf{x}}^{(\alpha)}\big\}; \underline{\mathbf{w}}\Big) = \prod_{\alpha=1}^{p} \widehat{P}\big(\underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}\big)$$

Model selection and Maximum Likelihood

$$\widehat{P}\left(\{\underline{\mathbf{x}}^{(\alpha)}\};\underline{\mathbf{w}}\right) \overset{!}{=} \max_{(\underline{\mathbf{w}})}$$

intuition: select the model which generates the observed data with high probability

in practice: minimization of the negative $\log$-likelihood

$$\begin{aligned} p \cdot E^T_{[\underline{\mathbf{w}}]} &= -\ln \widehat{P}\left(\{\underline{\mathbf{x}}^{(\alpha)}\};\underline{\mathbf{w}}\right) \\ &= -\sum_{\alpha=1}^{p} \ln \widehat{P}\left(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}}\right) \\ &\overset{!}{=} \min \end{aligned}$$

- equivalent to the minimization of the KL-divergence via ERM.

The multivariate Gaussian

$$\widehat{P}\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\};\underline{\boldsymbol{\mu}},\underline{\boldsymbol{\Sigma}}\right) = \left(\frac{1}{\sqrt{(2\pi)^N \det \underline{\boldsymbol{\Sigma}}}}\right)^p \cdot \prod_{\alpha=1}^{p} \exp\left(-\frac{1}{2}\left(\underline{\mathbf{x}}^{(\alpha)} - \underline{\boldsymbol{\mu}}\right)^T \underline{\boldsymbol{\Sigma}}^{-1}\left(\underline{\mathbf{x}}^{(\alpha)} - \underline{\boldsymbol{\mu}}\right)\right)$$

$$\begin{aligned} E^T\left(\underline{\boldsymbol{\mu}},\underline{\boldsymbol{\Sigma}}\right) &= -\ln \widehat{P}\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\};\underline{\boldsymbol{\mu}},\underline{\boldsymbol{\Sigma}}\right) \\ &= \frac{p \cdot N}{2}\ln(2\pi) + \frac{p}{2}\ln(\det \underline{\boldsymbol{\Sigma}}) + \frac{1}{2}\sum_{\alpha=1}^{p}\left(\underline{\mathbf{x}}^{(\alpha)} - \underline{\boldsymbol{\mu}}\right)^T \underline{\boldsymbol{\Sigma}}^{-1}\left(\underline{\mathbf{x}}^{(\alpha)} - \underline{\boldsymbol{\mu}}\right) \end{aligned}$$

minimization of E^T (necessary conditions):

$$\frac{\partial E^T}{\partial \underline{\boldsymbol{\mu}}} = \underline{\mathbf{0}} \quad \Rightarrow \quad \underline{\boldsymbol{\mu}}^* = \frac{1}{p}\sum_{\alpha=1}^{p}\underline{\mathbf{x}}^{(\alpha)} \qquad \text{(empirical average)}$$

$$\frac{\partial E^T}{\partial \underline{\boldsymbol{\Sigma}}} = \underline{\mathbf{0}} \quad \Rightarrow \quad \underline{\boldsymbol{\Sigma}}^* = \frac{1}{p}\sum_{\alpha=1}^{p}(\underline{\mathbf{x}}^{(\alpha)} - \underline{\boldsymbol{\mu}}^*)(\underline{\mathbf{x}}^{(\alpha)} - \underline{\boldsymbol{\mu}}^*)^T \qquad \text{(empirical covariance matrix)}$$

remark: $\underline{\boldsymbol{\mu}}^*$ is unbiased, but $\underline{\boldsymbol{\Sigma}}^*$ is a biased estimator (cf. section: 5.2 Estimation theory)