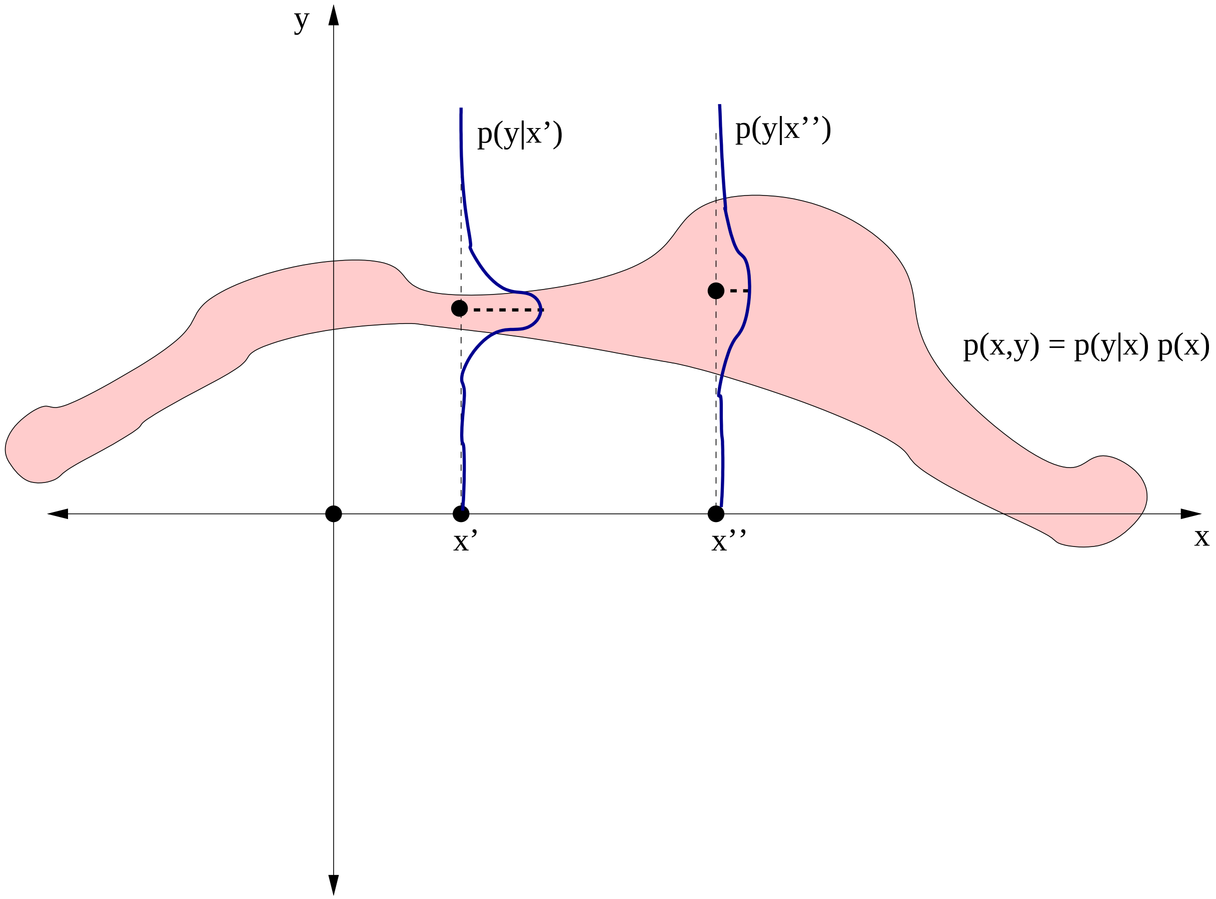
Intelligent Data Analysis

Density Modeling

Peter Tiňo
School of Computer Science
University of Birmingham

Supervised learning setting

E.g. Regression: need accurate model $p(y|\mathbf{x})$ of the conditional distribution of outputs y, given an input $\mathbf{x}$.



Input conditional distribution

Use normal distribution:

$$p(y|\mathbf{x}) \rightarrow N(\mu(\mathbf{x}), \sigma^2(\mathbf{x})),$$

or conditional 'ensemble' (mixture) of normal distributions

$$p(y|\mathbf{x}) = \sum_{j=1}^{M} P(j|\mathbf{x}) \cdot p(y|\mathbf{x}, j),$$

that is

$$p(y|\mathbf{x}) = \sum_{j=1}^{M} P(j|\mathbf{x}) \cdot \frac{1}{\sqrt{2\pi\sigma_j^2(\mathbf{x})}} \exp\left\{ -\frac{(y - \mu_j(\mathbf{x}))^2}{\sigma_j^2(\mathbf{x})} \right\}$$

Remember: $P(j|\mathbf{x}) \geq 0$, $\sum_j P(j|\mathbf{x}) = 1$, for every $\mathbf{x}$.

Representing the problem

We have 3 learning models cooperating with each other:

- $P(j|\mathbf{x})$
- $\mu_j(\mathbf{x}))$
- $\sigma_j^2(\mathbf{x})$

Can you construct a 'neural network like' structure to represent and this?

Collect all the model parameters in a parameter vector **w**: $p(y|\mathbf{x};\mathbf{w})$

Training via Maximum Likelihood

Given N training pairs $\mathcal{T} = \{(\mathbf{x}^1, y^1), (\mathbf{x}^2, y^2), ..., (\mathbf{x}^N, y^N)\}$, find parameter setting w_* that maximizes probability given by the model to the training sample:

$$\mathbf{w}_* = \underset{\mathbf{w}}{\operatorname{argmax}} \; p(\mathcal{T}|\mathbf{w})$$

Assume the example pairs are generated independently of each other:

$$p(\mathcal{T}|\mathbf{w}) = \prod_{i=1}^{N} p(y^i|\mathbf{x}^i; \mathbf{w}).$$

It is more convenient to maximize the log-likelihood

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} \log p(y^i|\mathbf{x}^i; \mathbf{w})$$

Example – Gaussians of the same variance

Assume a particularly simple model for the input-conditional distribution over outputs:

$$p(y|\mathbf{x};\mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \ \exp\left\{ -\frac{(y-\mu(\mathbf{x};\mathbf{w}))^2}{\sigma^2} \right\}$$

In this case,

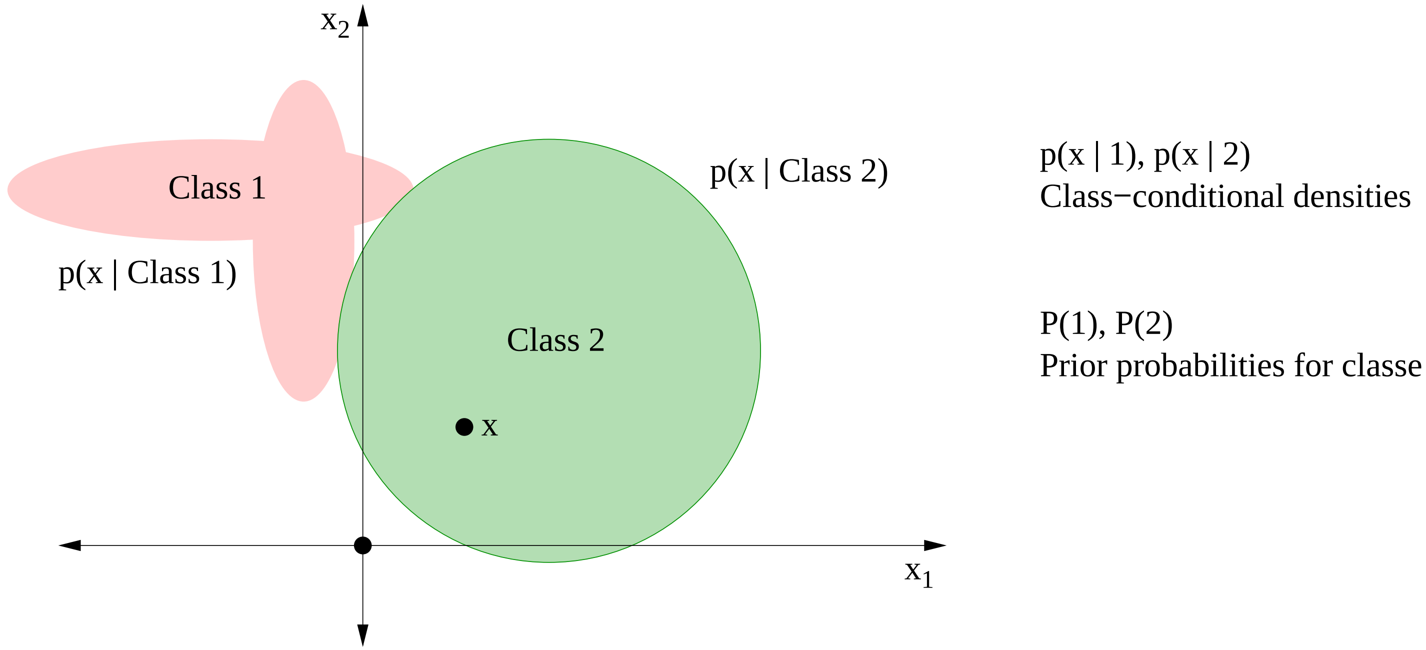
$$\operatorname*{argmax}_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \operatorname*{argmax}_{\mathbf{w}} \sum_{i=1}^{N} -(y^i - \mu(\mathbf{x}^i;\mathbf{w}))^2$$

Hence, optimal parameter setting $\mathbf{w}_*$ can be found by minimizing sum of squared errors

$$\mathcal{E}(\mathbf{w}) = \sum_{i=1}^{N} (y^i - \mu(\mathbf{x}^i;\mathbf{w}))^2$$

Unsupervised learning + Classification

Good density estimation can be cructial as a pre-processing step for other task, e.g. classification.



$$P(j|\mathbf{x}) = \frac{p(\mathbf{x}|j) \cdot P(j)}{p(\mathbf{x}|1) \cdot P(1) + p(\mathbf{x}|2) \cdot P(2)}, \quad j = 1, 2$$