6.891 Machine learning and neural networks

Tommi Jaakkola
MIT AI LAB
tommi@ai.mit.edu

LECTURE 10 (October 12, 2000)

Topics

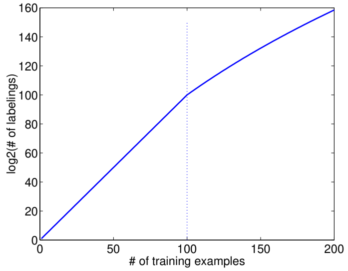
- Model selection
  - Basic idea
  - Structural risk minimization
  - Example
- Density estimation
  - Parametric, mixture models
  - Estimation via the EM algorithm
  - Non-parametric models

Learning and VC-dimension

- Terminology from previous lecture

1. A set of classifiers
2. Complexity
3. Shattering
4. VC-dimension

Learning and VC-dimension



- The number of labelings that the set of classifiers can generate over n points increases sub-exponentially after $n > d_{VC}$ (in this case $d_{VC} = 100$)

Learning and VC-dimension

- *Finite VC-dimension is a sufficient condition for (exponentially) fast convergence of a learning method*

  By convergence we mean:

  $$\overbrace{\frac{1}{n}\sum_{i=1}^{n} \text{Loss}(y_i, h(\mathbf{x}_i))}^{\text{Empirical loss}} - \overbrace{E\{\,\text{Loss}(y, h(\mathbf{x}))\,\}}^{\text{Expected loss}} \rightarrow 0$$

  uniformly for all $h \in F$. Here $\text{Loss}(y, h(\mathbf{x})) = 1$ if $y \neq h(\mathbf{x})$ and zero otherwise (so called zero-one loss)

- This result holds for **any** underlying probability distribution from which the examples and the labels are generated

Model selection

- Model selection concerns with trying to balance the complexity of the model with the fit to the training data
- We need to have a (preferably) nested sequence of models of increasing complexity

  Model 1 $d_{VC} = d_1$
  Model 2 $d_{VC} = d_2$
  Model 3 $d_{VC} = d_3$

  where $d_1 \leq d_2 \leq d_3 \leq \ldots$

- Basic formulation: we derive a model selection criterion:

  $$\text{Expected loss} \leq \text{Empirical loss} + \text{Complexity penalty} \qquad (1)$$

  We select the model with the lowest bound on the expected loss.

Model selection: example

- Models of increasing complexity

  Model 1 $K(\mathbf{x}_1, \mathbf{x}_2) = (1 + (\mathbf{x}_1^T \mathbf{x}_2))$
  Model 2 $K(\mathbf{x}_1, \mathbf{x}_2) = (1 + (\mathbf{x}_1^T \mathbf{x}_2))^2$
  Model 3 $K(\mathbf{x}_1, \mathbf{x}_2) = (1 + (\mathbf{x}_1^T \mathbf{x}_2))^3$
  ... ...

- These are nested, i.e.,

  $$F_1 \subseteq F_2 \subseteq F_3 \subseteq \ldots \qquad (2)$$

  where F_k refers to the set of possible decision boundaries that the model k can represent.

- Selection criterion?

Model selection: structural risk minimization

- Finite VC-dimension gives us some guarantees about how close the empirical loss is to the expected loss

  With probability at least $1 - \delta$ over the choice of the training set, for all $h \in F_k$

  $$\overbrace{E\{\,\text{Loss}(y, h(\mathbf{x}))\,\}}^{\text{Expected loss}} \leq \overbrace{\frac{1}{n}\sum_{i=1}^{n} \text{Loss}(y_i, h(\mathbf{x}_i))}^{\text{Empirical loss}} + \overbrace{\epsilon(n, \delta, d_k)}^{\text{Complexity penalty}} \qquad (3)$$

  where

  $$\begin{aligned} d_k &= \text{VC-dimension of model (set of hypothesis) } k \\ \delta &= \text{Confidence parameter (probability of failure)} \end{aligned}$$

- We find model k that has the lowest bound on the expected loss (right hand side of the above equation)

Structural risk minimization cont'd

- For our zero-one loss (classification error), we can derive the following complexity penalty (Vapnik 1995):

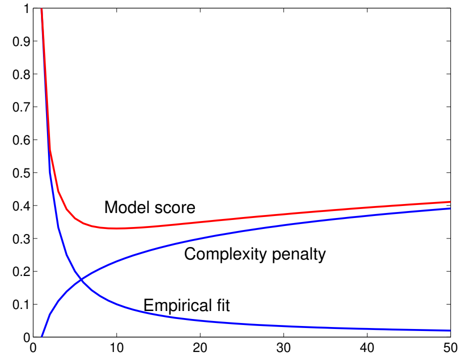
$$\epsilon(n,\delta,d) = \sqrt{\frac{d_{VC}(\log(2n/d_{VC})+1)+\log(1/(4\delta))}{n}} \qquad (4)$$

1. This is an increasing function of d_{VC}
2. Increases as δ decreases
3. Decreases as a function of n

(this is not the only choice...)

Structural risk minimization cont'd

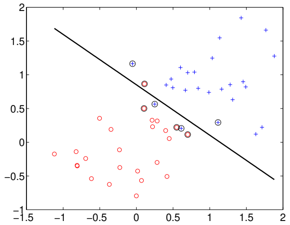
- Competition of terms...
  1. Empirical loss decreases with increasing d_{VC}
  2. Complexity penalty increases with increasing d_{VC}



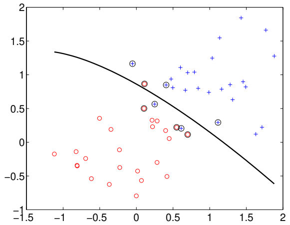
- We find the minimum of the combined score

Structural risk minimization: example

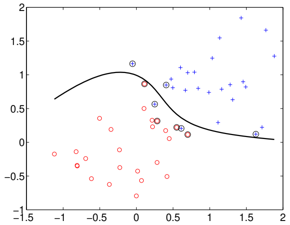
- The same problem as in the previous lecture



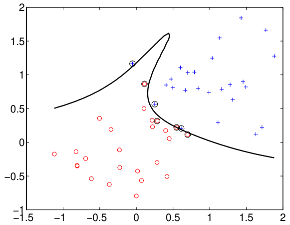
linear



2^{nd} order polynomial



4^{th} order polynomial



8^{th} order polynomial

Structural risk minimization: example cont'd

- Number of training examples $n = 50$, confidence parameter $\delta = 0.05$.

Model	d_{VC}	Empirical fit	Complexity penalty $\epsilon(n,\delta,d_{VC})$
1^{st} order	3	0.06	0.5501
2^{nd} order	6	0.06	0.6999
4^{th} order	15	0.04	0.9494
8^{th} order	45	0.02	1.2849

- Structural risk minimization would clearly select the simplest (linear) model in this case.

Where we are

- Model selection
  - Basic idea
  - Structural risk minimization
  - Example
- *Density estimation*
  - Parametric, mixture models
  - Estimation via the EM algorithm
  - Non-parametric models

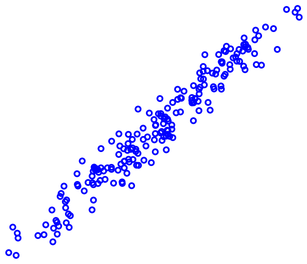
Parametric density models

- Probability model = a class of probability distributions
- Example: a simple multivariate Gaussian model

$$P(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{ -\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu) \} \quad (5)$$

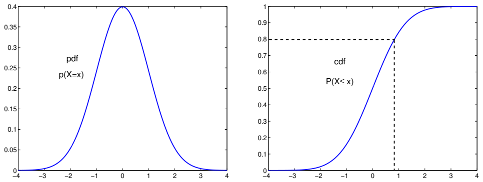
- This is a *generative model* in the sense that we can generate $\mathbf{x}$'s

  How do I generate a sample from a specific multivariate Gaussian distribution?



Gaussian samples

- 1-dimensional Gaussian *probability density function* (pdf) $P(x|\mu, \Sigma)$ and *cumulative distribution function* (cdf)



- To draw a sample from a Gaussian, we can invert the cumulative distribution function $F(x) = \int_{-\infty}^{x} P(z|\mu, \Sigma)dz$:

$$u \sim \text{Uniform}(0,1) \;\Rightarrow\; x = F^{-1}(u) \sim P(x|\mu, \Sigma) \quad (6)$$

- A multivariate sample can be constructed from multiple independent one dimensional Gaussian samples $\mathbf{z} = [z_1, \ldots, z_d]^T$:

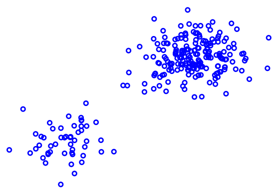
$$\mathbf{x} = \Sigma^{1/2}\mathbf{z} + \mu \;\Rightarrow\; \mathbf{x} \sim P(x|\mu, \Sigma) \quad (7)$$

Parametric density models

- Example: a mixture of Gaussian model

$$P(\mathbf{x}|\theta) = \sum_{i=1}^{k} p_j\, P(\mathbf{x}|\mu_j, \Sigma_j) \quad (8)$$

where $\theta = \{p_1, \ldots, p_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k\}$ contains all the parameters of the mixture model. $\{p_j\}$ are known as *mixing proportions or coefficients*.



- How do we generate a sample from the mixture?
- How do we estimate such a mixture model?

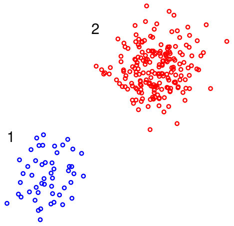
Mixture density estimation

(For simplicity, we'll look at only maximum likelihood estimation)

Suppose we want to estimate a two component mixture model.

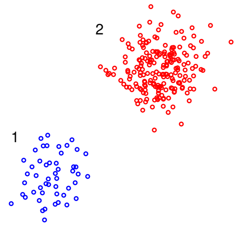
$$P(\mathbf{x}|\theta) = p_1\, P(\mathbf{x}|\mu_1, \Sigma_1) + p_2\, P(\mathbf{x}|\mu_2, \Sigma_2) \tag{9}$$

- If each example $\mathbf{x}_j$ in the training set were labeled $y_i = 1, 2$ according to which mixture component (1 or 2) generated it, then the estimation would be easy.



Mixture density estimation

If the examples were labeled, we could estimate each Gaussian independently of each other



- Separately for $l = 1, 2$

$$\hat{n}_l \leftarrow \sum_{i:y_i=l} 1 = \#\text{ of examples labeled } l \tag{10}$$

$$\hat{p}_l \leftarrow \frac{\hat{n}_l}{n} \tag{11}$$

$$\hat{\mu}_l \leftarrow \frac{1}{\hat{n}_l} \sum_{i:y_i=l} \mathbf{x}_i \tag{12}$$

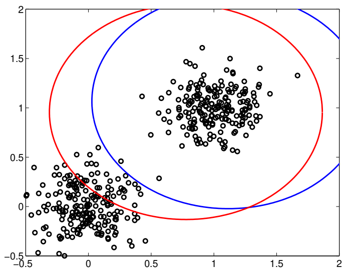
$$\hat{\Sigma}_l \leftarrow \frac{1}{\hat{n}_l} \sum_{i:y_i=l} (\mathbf{x}_i - \hat{\mu}_l)(\mathbf{x}_i - \hat{\mu}_l)^T \tag{13}$$

Mixture density estimation

- Of course we don't have such a labels ... but we can guess what the labels might be based on our current mixture distribution
- We get soft labels, posterior probabilities of which Gaussian generated which example:

$$\hat{p}(l|i) \leftarrow P(y_i = l|\mathbf{x}_i, \theta) \tag{14}$$

where $\sum_{l=1,2} \hat{p}(l|i) = 1$.



The EM algorithm

E-step: First we perform a soft reassignment of examples based on the current mixture distribution

$$\hat{p}(l|i) \leftarrow P(y_i = l|\mathbf{x}_i, \theta), \text{ for all } l = 1, 2 \text{ and } i = 1, \ldots, n \tag{15}$$

M-step: Then we re-estimate the parameters (separately for the two Gaussians) based on the soft assignments.

$$\hat{n}_l \leftarrow \sum_{i=1}^{n} \hat{p}(l|i) = \text{Soft } \#\text{ of examples labeled } l \tag{16}$$

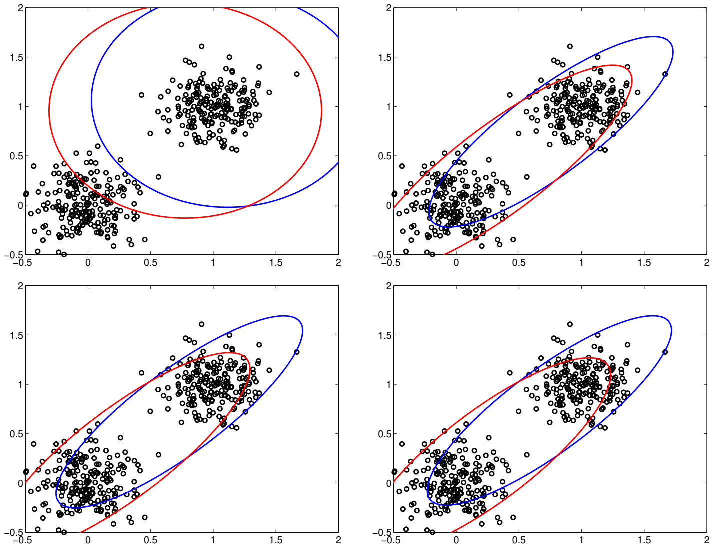
$$\hat{p}_l \leftarrow \frac{\hat{n}_l}{n} \tag{17}$$

$$\hat{\mu}_l \leftarrow \frac{1}{\hat{n}_l} \sum_{i=1}^{n} \hat{p}(l|i)\, \mathbf{x}_i \tag{18}$$

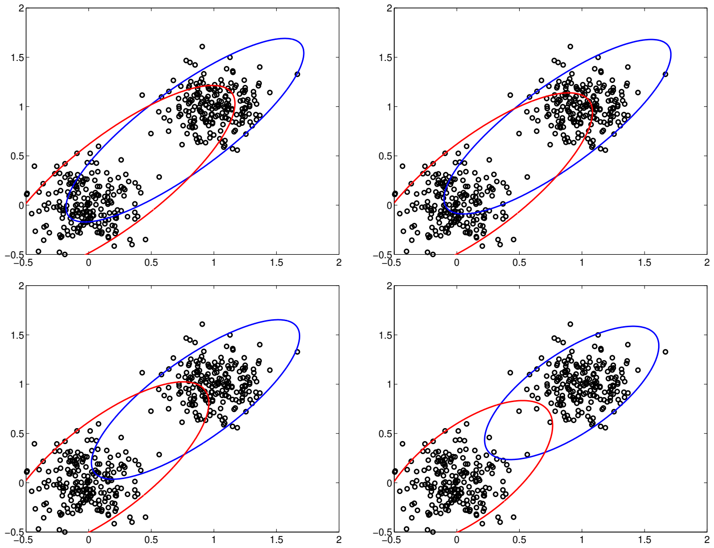
$$\hat{\Sigma}_l \leftarrow \frac{1}{\hat{n}_l} \sum_{i=1}^{n} \hat{p}(l|i)\, (\mathbf{x}_i - \hat{\mu}_l)(\mathbf{x}_i - \hat{\mu}_l)^T \tag{19}$$

where $l = 1, 2$.

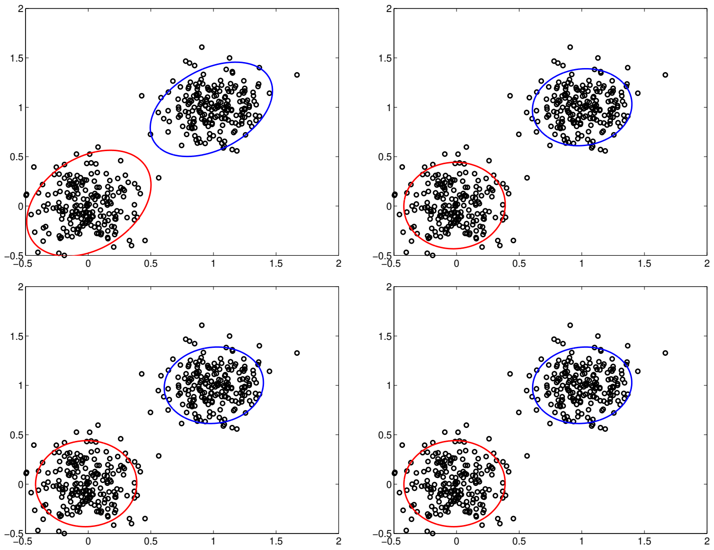
Mixture density estimation: example



Mixture density estimation

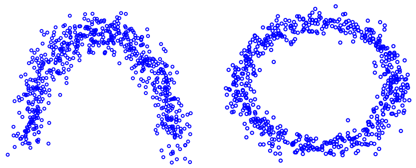


Mixture density estimation



Parametric density models cont'd

- More mixture densities



- We can approximate almost any distribution by including more and more components in the mixture model

$$P(\mathbf{x}|\theta) = \sum_{i=1}^{k} p_j \, P(\mathbf{x}|\mu_j, \Sigma_j) \tag{20}$$

Non-parametric densities

- We can assign one mixture component around each example (Parzen window):

$$\hat{P}_{parzen}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} P(\mathbf{x}|\mathbf{x}_j, \sigma^2 I) \tag{21}$$

where n is the number of examples.

- The covariance matrices for each bump $\Sigma_j = \sigma^2 \cdot I$ are all equal and isotropic. The single parameter σ^2 controls smoothness of the density esitmate