Machine Learning
Exponential Family

Dmitrij Schlesinger

WS2013/2014, 22.11.2013

COMPUTER VISION LAB DRESDEN

TECHNISCHE UNIVERSITÄT DRESDEN

General form

$$p(x;\theta) = h(x)\exp\Big[\langle\eta(\theta), T(x)\rangle - A(\theta)\Big]$$

with

- x is a random variable
- θ is a parameter
- $\eta(\theta)$ is a natural parameter, vector (often $\eta(\theta) = \theta$)
- $T(x)$ is a sufficient statistic
- $A(\theta)$ is the log-partition function

Almost all probability distributions you can imagine are members of the exponential family

Example – Gaussian (board)

Our models

Let x be an observed variable and y be a hidden one

1. The joint probability distribution is in the exponential family (a generative model):

$$p(x,y;w) = \frac{1}{Z(w)} \exp\Big[\langle \phi(x,y), w \rangle\Big]$$

$$Z(w) = \sum_{x,y} \exp\Big[\langle \phi(x,y), w \rangle\Big]$$

2. The conditional probability distribution is in the exponential family (a discriminative model):

$$p(x,y;w) = p(x) \cdot p(y|x;w)$$

$$p(y|x;w) = \frac{1}{Z(w,x)} \exp\Big[\langle \phi(x,y), w \rangle\Big]$$

$$Z(w,x) = \sum_{y} \exp\Big[\langle \phi(x,y), w \rangle\Big] \quad \forall x$$

Our learning schemes

- Generative model, supervised $\rightarrow$ Maximum Likelihood, Gradient
- Discriminative model, supervised $\rightarrow$ Maximum Conditional Likelihood, Gradient
- Generative model, unsupervised $\rightarrow$ Maximum Likelihood, Expectation Maximization, Gradient for the M-step

Generative model, supervised

Model:

$$p(x, y; w) = \frac{1}{Z(w)} \exp\Big[\langle\phi(x, y), w\rangle\Big]$$

$$Z(w) = \sum_{x,y} \exp\Big[\langle\phi(x, y), w\rangle\Big]$$

Training set: $L = \Big((x^l, y^l) \ldots\Big)$

Maximum Likelihood:

$$\sum_l \Big[\langle\phi(x^l, y^l), w\rangle - \ln Z(w)\Big] \to \min_w$$

Gradient:

$$\frac{\partial}{\partial w} = \frac{1}{|L|} \sum_l \phi(x^l, y^l) - \frac{\partial \ln Z(w)}{\partial w}$$

Generative model, supervised

Partition function:

$$Z(w) = \sum_{x,y} \exp\Big[\langle \phi(x,y), w \rangle\Big]$$

Gradient of the log-partition function:

$$\begin{aligned}
\frac{\partial \ln Z(w)}{\partial w} &= \\
&= \frac{1}{Z(w)} \sum_{x,y} \exp\Big[\langle \phi(x,y), w \rangle\Big] \cdot \phi(x,y) \\
&= \sum_{x,y} p(x,y;w) \cdot \phi(x,y) = \mathbb{E}_{p(x,y;w)}[\phi]
\end{aligned}$$

The Gradient is the difference of expectations:

$$\frac{\partial}{\partial w} = \frac{1}{|L|} \sum_{l} \phi(x^l, y^l) - \mathbb{E}_{p(x,y;w)}[\phi] = \mathbb{E}_L[\phi] - \mathbb{E}_{p(x,y;w)}[\phi]$$

Discriminative model (posterior), supervised

Model:

$$p(y|x;w) = \frac{1}{Z(w,x)} \exp\Big[\langle\phi(x,y),w\rangle\Big]$$

$$Z(w,x) = \sum_y \exp\Big[\langle\phi(x,y),w\rangle\Big] \quad \forall x$$

Training set: $L = \Big((x^l,y^l)\ldots\Big)$

Maximum Conditional Likelihood:

$$\sum_l \Big[\langle\phi(x^l,y^l),w\rangle - \ln Z(w,x^l)\Big] \to \min_w$$

Gradient:

$$\frac{\partial}{\partial w} = \frac{1}{|L|}\sum_l \phi(x^l,y^l) - \frac{1}{|L|}\sum_l \frac{\partial \ln Z(w,x^l)}{\partial w}$$

Discriminative model (posterior), supervised

Partition function:

$$Z(w,x) = \sum_y \exp\Big[\langle \phi(x,y), w\rangle\Big]$$

Gradient of the log-partition function for a particular x^l:

$$\begin{aligned}\frac{\partial \ln Z(w,x^l)}{\partial w} &= \\ &= \frac{1}{Z(w,x^l)} \sum_y \exp\Big[\langle \phi(x^l,y), w\rangle\Big] \cdot \phi(x^l,y) \\ &= \sum_y p(y|x^l;w) \cdot \phi(x^l,y) = \mathbb{E}_{p(y|x^l;w)}\Big[\phi(x^l)\Big]\end{aligned}$$

The Gradient is again the difference of expectations:

$$\frac{\partial}{\partial w} = \mathbb{E}_L[\phi] - \frac{1}{|L|} \sum_l \mathbb{E}_{p(y|x^l;w)}\Big[\phi(x^l)\Big]$$

Generative model, unsupervised

Model:

$$p(x,y;w) = \frac{1}{Z(w)} \exp\Big[\langle \phi(x,y), w\rangle\Big]$$

$$Z(w) = \sum_{x,y} \exp\Big[\langle \phi(x,y), w\rangle\Big]$$

Training set (incomplete): $L = \left(x^l \ldots\right)$

Expectation:

$$\alpha_l(y) = p(y|x^l;w) \quad \forall l, y$$

Maximization:

$$\sum_l \sum_y \alpha_l(y) \ln p(x,y;w) \to \max_w$$

Generative model, unsupervised

Maximization:

$$\sum_l \sum_y \alpha_l(y) \ln p(x,y;w) =$$
$$= \sum_l \sum_y \alpha_l(y) \Big[\langle \phi(x^l,y), w\rangle - \ln Z(w) \Big] =$$
$$= \sum_l \sum_y \alpha_l(y) \langle \phi(x^l,y), w\rangle - \sum_l \sum_y \alpha_l(y) \ln Z(w) =$$
$$= \sum_l \sum_y \alpha_l(y) \langle \phi(x^l,y), w\rangle - |L| \cdot \ln Z(w)$$

The gradient is again a difference of expectations:

$$\frac{\partial}{\partial w} = \frac{1}{|L|} \sum_l \sum_y \alpha_l(y) \phi(x^l,y) - \mathbb{E}_{p(x,y;w)}[\phi] =$$
$$= \frac{1}{|L|} \sum_l \mathbb{E}_{p(y|x^l)}[\phi(x^l)] - \mathbb{E}_{p(x,y;w)}[\phi]$$

Conclusion

In all variants the gradient of the log-likelihood is a difference between expectations of the sufficient statistic:

$$\boxed{\frac{\partial \ln L}{\partial w} = \mathbb{E}_{data}[\phi] - \mathbb{E}_{model}[\phi]}$$

$\rightarrow$ the likelihood is in optimum if they coincide

In supervised cases the "data" expectation is the simple average over the training set $\rightarrow$ $\mathbb{E}_{data}$ does not depend on w $\rightarrow$ the problem is concave $\rightarrow$ global optimum.