Deep learning.

- Model parametrization
- Defining a loss function
- Optimization
- Model selection
- Reporting

Parametrization

$$p(\underbrace{y}_{\text{output}}|\underbrace{x}_{\text{input}})$$

1) $p(y|x) \propto \exp\{F(x,y)\}$

2) latent variable approach

$$p(y,z|x) \propto \exp\{F(x,y,z)\}$$

$$p(z|x) \propto \exp\{F'(x,z)\}$$

Multiclass classification.

$y \in \{1, \cdots, c\}$

$$F(x,y) = \underset{\mathbb{R}^d}{\underset{\uparrow}{w_y^T}} \, \underset{\mathbb{R}^d}{\underset{\uparrow}{\phi(x)}} + \underset{\mathbb{R}}{\underset{\uparrow}{b_y}}$$

$$p(y|x) = \frac{\exp\{w_y^T \phi(x) + b_y\}}{\sum_{y'=1}^{c} \exp\{w_{y'}^T \phi(x) + b_y\}}$$

Mixture density network

$y \in \mathbb{R}^{d'}$

$z \in \{1, \cdots, c\}$

$$F(x,y,z) = -\| \overset{\in \mathbb{R}^{d'}}{y} - \overset{\mathbb{R}^{d' \times d}}{\overset{\swarrow}{W_\mu^z}} \overset{\in \mathbb{R}^d}{\phi(x)} \|^2$$

$$F'(x,z) = \underset{\mathbb{R}^d}{\underset{\uparrow}{w_z^T}} \phi(x) + \underset{\mathbb{R}}{\underset{\uparrow}{b_z}}$$

$$p(y|x) = \sum_{z=1}^{c} p(y,z|x) = \sum_{z=1}^{c} p(y|z,x) \underbrace{p(z|x)}$$

$$= \sum_{z=1}^{c} \left(\frac{\exp\{w_z^T \phi(x) + b_z)}{\sum_{z'=1}^{c} \exp\{w_{z'}^T \phi(x) + b_{z'}\}} \right) \left(\frac{\exp\{-\| y - W_\mu^z \phi(x) \|^2\}}{\int_{\mathbb{R}^d} \exp\{-\| y' - W_\mu^z \phi(x) \|^2\} dy'} \right).$$

Loss function

for (x, y^*), $\overset{\swarrow \text{ per-example loss}}{l}$ is $-\log p(y^*|x)$

for $D = \{(x_1, y_1^*), \cdots, (x_N, y_N^*)\}$,

$$p(D) = \prod_{n=1}^{N} p(x_n, y_n^*) = \prod_{n=1}^{N} p(y_n^*|x_n) \underbrace{p(x_n)}_{\text{uniform}}$$

$$\log p(D) = \sum_{n=1}^{N} \log p(y_n^*|x_n) + \text{const.}$$

Multiway classification

$$-\log p(y^*|x) = \underbrace{-w_{y^*}^T \phi(x) - b_{y^*}}_{= F(x,y)} + \log \sum_{y'=1}^{c} \exp\{w_{y'}^T \phi(x) + b_{y'}\}$$

Optimization : Gradient descent

$L(D;\theta)$: a data-level loss fn.

$$L(D;\theta) = \frac{1}{|D|} \sum_{(x,y)\in D} l(x,y;\theta)$$

$$\theta \leftarrow \theta - \eta \underbrace{\overset{\swarrow \text{gradient}}{\nabla_\theta L(D;\theta)}}_{} = \theta - \underbrace{\frac{\eta}{|D|} \sum_{x,y\in D} \nabla_\theta l(x,y;\theta)}_{}$$

for $(x,y) \in D$, $x,y \sim P_D$

$\overset{\swarrow \text{minibatch}}{M} \subset D \quad |M| << |D|$

$$\nabla_\theta L(D;\theta) \approx \underline{\mathbb{E}_{x,y\sim P_D}[\nabla l(x,y;\theta)]} \approx \underline{\frac{1}{|M|} \sum_{(x,y)\in M} \nabla l(x,y;\theta)}$$

Stochastic Gradient Descent

Model selection

= hyperparameter selection/optimization

$\lambda \in \Lambda$: a hyperparameter

$$\min_{\lambda\in\Lambda} \underset{\substack{\uparrow \\ \text{validation set } D'}}{\mathbb{E}_{(x,y)\sim D}} \left[l'\left(x,y ; \arg\min_\theta \underset{\substack{\uparrow \\ \text{training set } D}}{\mathbb{E}_{(x,y)\sim D}}[l(x,y;\theta)]\right)\right]$$

Reporting