TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2018

Connectionist Temporal Classification (CTC)

A Latent Variable Deep Graphical Model

The General Expectation Gradient (EG) Algorithm

The Big Picture I

Conditional vs. Unconditional

$$\Phi^* = \operatorname*{argmin}_{\Phi} E_{(x,y)\sim \mathrm{Pop}}\ -\ln\ P(y|x)$$

$$\Phi^* = \operatorname*{argmin}_{\Phi} E_{y\sim \mathrm{Pop}}\ -\ln\ P(y)$$

This is a non-distinction: the issues in the to the conditional case are exactly the same as in the unconditional case.

The Big Picture II

The binary case: $y \in \{-1, 1\}$ (cancer screening).

The multiclass case: $y \in \mathcal{Y}$ where iteration over $\hat{y} \in \mathcal{Y}$ is feasible (MNIST, CFAR, ImageNet).

The structured case: $y \in \mathcal{Y}$ where $\mathcal{Y}$ is discrete but iteration over $\hat{y} \in \mathcal{Y}$ is infeasible (language modeling, speech recognition).

Big Picuture III

Graphical models (such as CTC) rely on assumptions about the structure of $P_\Phi(y)$.

The expectation gradient (EG) algorithm applies when $P_\Phi(y)$ can be computed exactly using dynamic programming.

Requiring that $P_\Phi(y)$ is computable restricts the model but is justified in some cases (such as CTC).

Connectionist Temporal Classification (CTC)

Phonetic Transcription

A speech signal

$$x = x_1, \ \ldots, \ x_T$$

is labeled with a phone sequence

$$y = y_1, \ldots, y_N$$

with $N << T$ and with $y_n \in \mathcal{Y}$ for a set of phonemes $\mathcal{Y}$.

The length N of y is not determined by x and the alignment between x and y is not given.

CTC

The model defines $P_\Phi(\hat{z}|x)$ where $\hat{z}$ is latent and where y is determined by $\hat{z}$.

$$\hat{z} = \hat{z}_1, \ \ldots, \ \hat{z}_T, \quad \hat{z}_t \in \mathcal{Y} \cup \{\perp\}$$

The sequence

$$y(\hat{z}) = y_1, \ \ldots, \ y_N$$

is the result of removing all the occurrences of $\perp$ from $\hat{z}$.

$$\perp, a_1, \perp, \perp, \perp, a_2, \perp, \perp, a_3, \perp \Rightarrow a_1, a_2, a_3$$

The CTC Model

$$h_1,\ \ldots,\ h_T = \mathrm{RNN}_\Phi(x_1,\ \ldots,\ x_T)$$

$$P_\Phi(\hat{z}_t|x_1,\ldots,x_T) = \operatorname*{softmax}_{\hat{z}_t}\ e(\hat{z}_t)^\top h_t$$

Where $e(\hat{z}_t)$ is a vector embedding of the phoneme $\hat{z}_t$. The embedding is a parameter of the model.

Note that $\hat{z}_1, \ldots\ \hat{z}_T$ are all independent given x.

The Expectation Gradient (EG) Algorithm

CTC is a special case of a latent variable graphical model.

In cases where $P_\Phi(y|x)$ can be computed from dynamic programming we can backpropagate through the dynamic programming algorithm to get the gradient of cross-entropy loss.

This general technique is the **expectation gradient** (EG) algorithm.

We will first show that for the CTC model we can compute $P_\Phi(y|x)$ by dynamic programming.

We will then note that it is not necessary to backpropagate through the dynamic programming algorithm.

Dynamic Programming (Forward-Backward)

$$
\begin{aligned}
x &= x_1, \ \ldots, \ x_T \\
\hat{z} &= \hat{z}_1, \ \ldots, \ \hat{z}_T, \quad \hat{z}_t \in \mathcal{Y} \cup \{\perp\} \\
y &= y_1, \ \ldots, \ y_N, \quad y_n \in \mathcal{Y}, \quad N << T \\
y(\hat{z}) &= (\hat{z}_1, \ \ldots, \ \hat{z}_T) - \perp
\end{aligned}
$$

Forward-Backward

$$
\begin{aligned}
\vec{y}_t &= (\hat{z}_1, \ \ldots, \ \hat{z}_t) - \perp \\
F[n,t] &= P(\vec{y}_t = y_1, \ \ldots, \ y_n) \\
B[n,t] &= P(y_{n+1}, \ \ldots, y_N | \vec{y}_t = y_1, \ \ldots, \ y_n) \\
P(y) &= F[N,T] = B[0,0]
\end{aligned}
$$

Dynamic Programming (Forward-Backward)

$$\begin{aligned}
\vec{y}_t &= (\hat{z}_1,\ \ldots,\ \hat{z}_t) - \perp \\
F[n,t] &= P(\vec{y}_t = y_1,\ \ldots,\ y_n) \\
B[n,t] &= P(y_{n+1},\ \ldots, y_N | \vec{y}_t = y_1,\ \ldots,\ y_n)
\end{aligned}$$

$$\begin{aligned}
F[0,0] &= 1 \\
F[n,0] &= 0 \quad \text{for } n > 0 \\
F[n+1,t+1] &= P(\hat{z}_{t+1} = \perp)F[n+1,t] + P(\hat{z}_{t+1} = y_{n+1})F[n,t]
\end{aligned}$$

$$\begin{aligned}
B[N,T] &= 1 \\
B[n,T] &= 0, \quad \text{for } n < N \\
B[n-1,t-1] &= P(\hat{z}_t = \perp)B[n-1,t] + P(\hat{z}_t = y_n)B[n,t]
\end{aligned}$$

The Big Picture I

Conditional vs. Unconditional

$$\Phi^* = \operatorname*{argmin}_{\Phi} E_{(x,y)\sim\mathrm{Pop}}\ -\ln\ P(y|x)$$

$$\Phi^* = \operatorname*{argmin}_{\Phi} E_{y\sim\mathrm{Pop}}\ -\ln\ P(y)$$

This is a non-distinction: the issues in the to the conditional case are exactly the same as in the unconditional case.

The Big Picture II

The binary case: $y \in \{-1, 1\}$ (cancer screening).

The multiclass case: $y \in \mathcal{Y}$ where iteration over $\hat{y} \in \mathcal{Y}$ is feasible (MNIST, CFAR, ImageNet).

The structured case: $y \in \mathcal{Y}$ where $\mathcal{Y}$ is discrete but iteration over $\hat{y} \in \mathcal{Y}$ is infeasible (language modeling, speech recognition).

Graphical models (such as CTC) rely on assumptions about the structure of $P_\Phi(y)$.

END