1 Basics

1.1 Probability

Bayes rule: $p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y')p(y')dy'}$

$p(y|x)$ **posterior**, $p(x|y)$ **likelihood**, $p(y)$ **prior**, $p(x)$ or $\int p(x|y')p(y')dy'$ **marginal**

1.2 Activation Functions

ReLU(x): $\max(x, 0)$

ReLU'(x): $\begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$

Leaky ReLU(x): $\begin{cases} x & \text{if } x \geq 0 \\ 0.01x & \text{if } x < 0 \end{cases}$

Leaky ReLU'(x): $\begin{cases} 1 & \text{if } x \geq 0 \\ 0.01 & \text{if } x < 0 \end{cases}$

Sigmoid: $\sigma(x) = \frac{1}{1+\exp(-x)}$

$\sigma'(x) = \sigma(x)(1-\sigma(x))$

Hyperbolic tangent: $\tanh(x) = \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$

$\tanh'(x) = 1 - \tanh(x)^2$

2 Semirings

Semiring	Set	$\oplus$	$\otimes$	$\bar{0}$	$\bar{1}$	intuition/application
Boolean	$\{0,1\}$	$\vee$	$\wedge$	0	1	logical deduction, recognition
Viterbi	$[0,1]$	max	$\times$	0	1	prob. of the best derivation
Inside	$\mathbb{R}^+ \cup \{+\infty\}$	$+$	$\times$	0	1	prob. of a string
Real	$\mathbb{R} \cup \{+\infty\}$	min	$+$	$+\infty$	0	shortest-distance
Tropical	$\mathbb{R}^+ \cup \{+\infty\}$	min	$+$	$+\infty$	0	with non-negative weights
Counting	$\mathbb{N}$	$+$	$\times$	0	1	number of paths

- $(R, +)$ is a commutative monoid with identity element 0:
  - $(a+b)+c = a+(b+c)$
  - $0+a = a = a+0$
  - $a+b = b+a$
- $(R, \cdot)$ is a monoid with identity element 1:
  - $(a \cdot b) \cdot c = a \cdot (b \cdot c)$
  - $1 \cdot a = a = a \cdot 1$
- Multiplication left and right distributes over addition:
  - $a \cdot (b+c) = (a \cdot b) + (a \cdot c)$
  - $(a+b) \cdot c = (a \cdot c) + (b \cdot c)$
- Multiplication by 0 annihilates R:
  - $0 \cdot a = 0 = a \cdot 0$

3 Log-Linear Models

Log-lin. model: $p(y|x,\theta) = \frac{1}{Z(x,\theta)} \exp(\theta f(x,y))$

$\log(p(y|x,\theta)) = \theta f(x,y) + \text{const.}$

where $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^K$ is the feature function $Z(\theta) = \sum_{y' \in \mathcal{Y}} \exp(\theta f(x,y'))$ is the partition function.

MLE: $\mathcal{L}(\theta) = \sum_{n=1}^{N} \log p(y_n|x_n,\theta)$

$\theta_{\text{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta)$

where Θ is compact subset of $\mathbb{R}^K$

$\nabla \mathcal{L}(\theta) = \sum_{i=1}^{N} f(x_n, y_n) - \sum_{n=1}^{N} \mathbb{E}_{Y \sim p(\cdot|x_n,\theta)}[f(x_n, Y$

$\sum_{n \leq N} f(x_n, y_n) = \sum_{n=1}^{N} \mathbb{E}_{Y \sim p(\cdot|x_n,\theta)}[f(x_n, Y)]$

4 Viterbi

$\mathbf{t}^* = \arg\max_{t \in \mathcal{T}^N} \exp(\text{score}(\mathbf{t}, \mathbf{w})) = \arg\max_{t \in \mathcal{T}^N} \prod_{n=1}^{N} \exp\{\text{score}(\langle t_{n-1}, t_n \rangle, w)\}$

Algorithm 5.2

def *Viterbi Algorithm*$(\mathbf{w}, \mathcal{T}, N)$:
  for $t_{N-1} \in \mathcal{T}$:
    $v(\mathbf{w}, t_{N-1}, N-1) \leftarrow \exp(\text{score}(\langle t_{N-1}, \textsc{eos} \rangle, \mathbf{w}))$
  end for
  for $n \in N-2, \dots, 1$:
    for $t_n \in \mathcal{T}$:
      $v(\mathbf{w}, t_n, n) \leftarrow \max_{t_{n+1} \in \mathcal{T}} \exp(\text{score}(\langle t_n, t_{n+1} \rangle, \mathbf{w})) \times v(\mathbf{w}, t_{n+1}, n+1)$
      $b(t_n, n) \leftarrow \text{argmax}_{t_{n+1} \in \mathcal{T}} \exp(\text{score}(\langle t_n, t_{n+1} \rangle, \mathbf{w})) \times v(\mathbf{w}, t_{n+1}, n+1)$ ▷ *Keeps track of the best tags*
    end for
  end for
  $v(\mathbf{w}, \textsc{bos}, 0) \leftarrow \max_{t_1 \in \mathcal{T}}(v(\mathbf{w}, \textsc{bos}, 0), \exp(\text{score}(\langle \textsc{bos}, t_1 \rangle, \mathbf{w})) \times v(\mathbf{w}, t_1, 1))$
  $b(\textsc{bos}, 0) \leftarrow \text{argmax}_{t_1 \in \mathcal{T}}(v(\mathbf{w}, \textsc{bos}, 0), \exp(\text{score}(\langle \textsc{bos}, t_1 \rangle, \mathbf{w})) \times v(\mathbf{w}, t_1, 1))$
  for $n \in 1, \dots, N$:
    $t_n \leftarrow b(t_{n-1}, n-1)$ ▷ *Recovers the best tagging sequence using backpointers*
  end for
  return $\mathbf{t}_{1:N}$, $v(\mathbf{w}, \textsc{bos}, 0)$

Runtime: $\mathcal{O}(N|\mathcal{T}|^2)$, $\mathcal{O}(N|\mathcal{T}|^3)$ when considering triplets.

5 Grammar

Definition 6.2

A ***context-free grammar*** *G is a quadruple $\langle \mathcal{N}, S, \Sigma, \mathcal{R} \rangle$ consisting of:*

- *A finite set of non-terminal symbols $\mathcal{N}$; written in upper-case letters, e.g. N_1, N_2, N_3*
- *A distinguished start non-terminal S*
- *An alphabet of terminal symbols Σ; written as lower-case letters, e.g. a_1, a_2, a_3*
- *A set of production rules $\mathcal{R}$ of the form $N \to \alpha$, where $N \in \mathcal{N}$ and $\alpha \in (\mathcal{N} \cup \Sigma)^*$ (Kleene closure of $\mathcal{N} \cup \Sigma$)*

5.1 Chomsky Normal Form

Grammar is in Chomsky Normal Form (CNF) if RHS of every production rule includes either two non-terminals or a single terminal symbol: $N_1 \to N_2 N_3$ or $N \to a$

5.2 PCFG

Probabilistic Context Free Grammar (PCFG) $\langle \mathcal{N}, S, \sum, \mathcal{R}, \mathcal{P} \rangle$, where $\mathcal{P}$ are probabilities assigned to each production rule.

5.3 WCFG

Weighted Context Free Grammar (WCFG) $\langle \mathcal{N}, S, \sum, \mathcal{R}, \mathcal{W} \rangle$, where $\mathcal{W}$ are non-negative weights assigned to each production rule. PCFG is special case of WCFG.

6 CKY

Algorithm 6.1

def *WeightedCKY*$(\mathbf{s}, \langle \mathcal{N}, S, \Sigma, \mathcal{R} \rangle, \text{score})$:
  $N \leftarrow |\mathbf{s}|$
  $chart \leftarrow 0$
  for $n = 1, \dots, N$:
    for $X \to \mathbf{s}_n \in \mathcal{R}$:
      $\text{chart}[n, n+1, X] \mathrel{+}= \exp\{\text{score}(X \to \mathbf{s}_n)\}$ ▷ *Handles single word tokens*
    end for
  end for
  for $span = 2, \dots, N$:
    for $i = 1, \dots, N - span + 1$: ▷ *i marks the beginning of the span*
      $k \leftarrow i + span$ ▷ *k marks the end of the span*
      for $j = i+1, \dots, k-1$: ▷ *j marks the breaking point of the span*
        for $X \to Y\,Z \in \mathcal{R}$:
          $\text{chart}[i, k, X] \mathrel{+}= \exp\{\text{score}(X \to Y\,Z)\} \times \text{chart}[i, j, Y] \times \text{chart}[j, k, Z]$
        end for
      end for
    end for
  end for
  return $\text{chart}[1, N+1, S]$