CSCI B609:
“Foundations of Data Science”

Lecture 10/11: Random Walks and Markov Chains + ML Intro

Slides at http://grigory.us/data-science-class.html

Grigory Yaroslavtsev

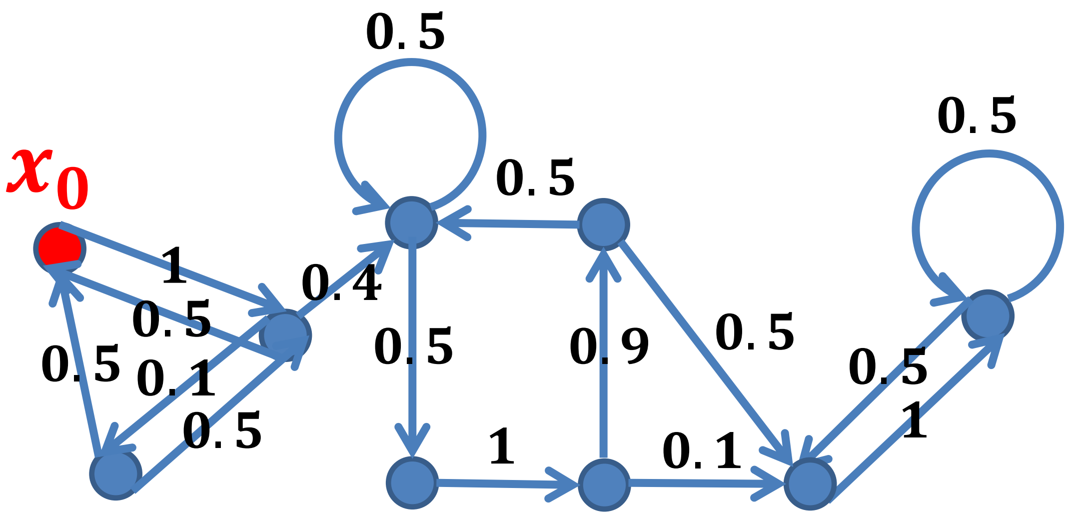
http://grigory.us

Project Example:
Gradient Descent in TensorFlow

- Gradient Descent (will be covered in class)
- Adagrad: http://www.magicbroom.info/Papers/DuchiHaSi10.pdf
- Momentum (stochastic gradient descent + tweaks): http://www.cs.toronto.edu/~hinton/absps/naturebp.pdf
- Adam (Adaptive + momentum): http://arxiv.org/pdf/1412.6980.pdf
- FTRL: http://jmlr.org/proceedings/papers/v15/mcmahan11b/mcmahan11b.pdf
- RMSProp: http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

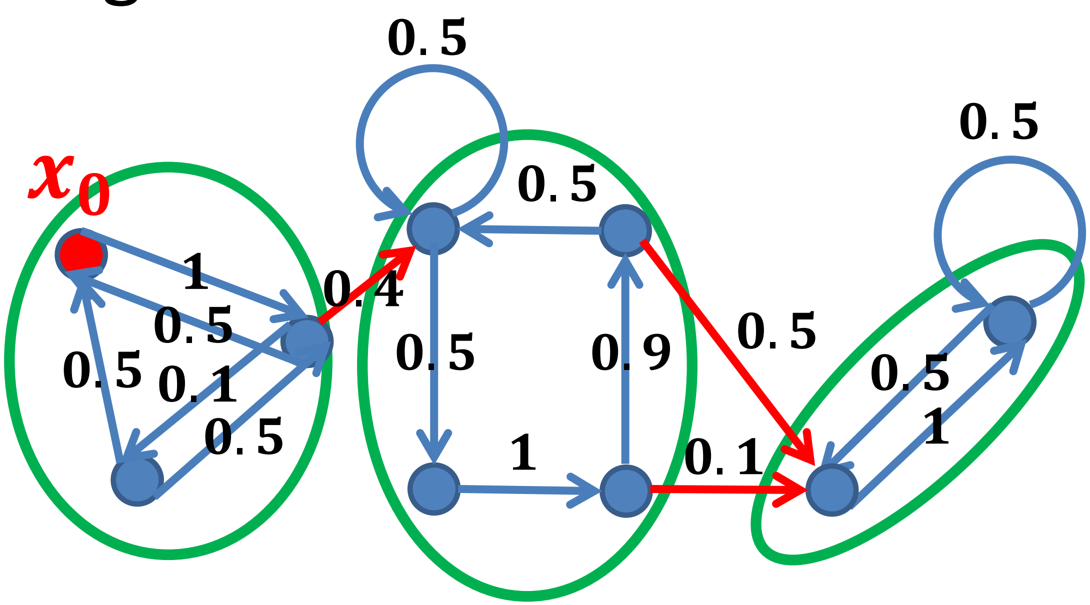
Random Walks and Markov Chains

- Random walk:
  - Directed graph $\boldsymbol{G}(\boldsymbol{V}, \boldsymbol{E})$
  - Starting vertex $\boldsymbol{x_0} \in \boldsymbol{V}$
  - Edge (i, j): probability p_{ij} of transition $i \rightarrow j$
  - $\forall i: \sum_j p_{ij} = 1$



Strongly Connected Components

- **Def (Strongly Connected Component).** $S \subseteq V$ such that $\forall i, j \in S$ there exist paths $i \to j$ and $j \to i$
- SCC's form a partition of the vertex set
- **Terminal SCC**: no outgoing edges
- Long enough random walk $\to$ **Terminal SCC**



Matrix Form and Stationary Distribution

- $\boldsymbol{p_t}$ = probability distribution over vertices at time $\boldsymbol{t}$
- $\boldsymbol{p}_0 = (1,0,0,\dots,0)$
- $\boldsymbol{p_t} P = \boldsymbol{p_{t+1}}$
- P = transition matrix with entries p_{ij}
- If $\boldsymbol{t} \to \infty$ then average of $\boldsymbol{p}_i's$ converges:

$$\frac{1}{\boldsymbol{t}} \sum_{\boldsymbol{i}=0}^{\boldsymbol{t-1}} \boldsymbol{p_i} \to \boldsymbol{\pi}$$

- $\boldsymbol{\pi}$ = **stationary distribution** of P
- $\boldsymbol{\pi}$ is unique and doesn't depend on x_0 if G is strongly connected
- Note: $\boldsymbol{p_t}$ for $\boldsymbol{t} \to \infty$ doesn't always converge!

Stationary Distribution

- Long-term average:

$$a_{\boldsymbol{t}} = \frac{1}{\boldsymbol{t}} \sum_{\boldsymbol{i}=0}^{\boldsymbol{t-1}} \boldsymbol{p_i}$$

- **Thm.** If G is strongly connected then $a_{\boldsymbol{t}} \rightarrow \boldsymbol{\pi}$:
  - $\boldsymbol{\pi} P = \boldsymbol{\pi}$
  - $\sum_i \boldsymbol{\pi}_i = 1$
  - $\boldsymbol{\pi}[P - I, \mathbf{1}] = [\mathbf{0}, 1]$
- We will show that $[P - I, \mathbf{1}]$ has rank $n \Rightarrow$ there is a unique solution to $\boldsymbol{\pi}[P - I, \mathbf{1}] = [\mathbf{0}, 1]$

Stationary Distribution Theorem

- **Thm.** $n \times (n+1)$ matrix $[P-I, \mathbf{1}]$ has rank n
- $A = [P-I, \mathbf{1}]$
- *Rank(A)* $< n \Rightarrow$ two lin. indep. solutions to $A\boldsymbol{x}=0$
- $\sum_j p_{ij} = 1 \Rightarrow \sum_j p_{ij} - 1 = 0$ (row sums of A)
  - $(\mathbf{1}, 0)$ is a solution to $A\boldsymbol{x} = 0$
- Assume there is another solution $(\boldsymbol{x}, \boldsymbol{\alpha}) \perp (\mathbf{1}, 0)$
  - $(P-I)\boldsymbol{x} + \alpha\mathbf{1} = \mathbf{0}$
  - $\forall i\colon \sum_j p_{ij}x_j - x_i + \alpha = 0 \Rightarrow x_i = \sum_j p_{ij}x_j + \alpha$
- $(\boldsymbol{x}, \boldsymbol{\alpha}) \perp (\mathbf{1}, 0) \Rightarrow$ not all $\boldsymbol{x}_j$ are equal

Stationary Distribution Theorem Cont.

- $\forall i\colon\ x_i = \sum_j p_{ij} x_j + \boldsymbol{\alpha}$
- $(\boldsymbol{x}, \alpha) \perp (\mathbf{1}, 0) \Rightarrow$ not all $\boldsymbol{x}_j$ are equal
- $\boldsymbol{S} = \{i : x_i = Max_{j=1}^{n} x_j\}$ = set of max value coord.
  - $\bar{\boldsymbol{S}}$ is non-empty
- G strongly connected $\Rightarrow \exists\ edge\ (k, l)\colon k \in \boldsymbol{S}, l \in \bar{\boldsymbol{S}}$
- $\Rightarrow x_k > \sum_j p_{kj} x_j \Rightarrow \boldsymbol{\alpha} > 0$
- Symmetric argument with $\boldsymbol{S} = \{i : x_i = Min_{j=1}^{n} x_j\}$
- $\Rightarrow x_{k'} < \sum_j p_{k'j} x_j \Rightarrow \boldsymbol{\alpha} < 0$
- Contradiction so $(\mathbf{1}, 0)$ is the unique solution

Fundamental Theorem of Markov Chains

- **Thm.** If P is transition matrix of a strongly connected Markov Chain and $a_t = \frac{1}{t}\sum_{i=0}^{t-1} \boldsymbol{p}_i$:
  - There exists a unique $\boldsymbol{\pi}$: $\boldsymbol{\pi}P = \boldsymbol{\pi}$
  - For any starting distribution: $\exists \lim_{t\to\infty} a_t = \boldsymbol{\pi}$
- a_t is a probability vector
- After one step: $a_t \to a_t P$
- $a_t P - a_t = \frac{1}{t}\left[\sum_{i=0}^{t-1} \boldsymbol{p}_i P\right] - \frac{1}{t}\left[\sum_{i=0}^{t-1} \boldsymbol{p}_i\right] = \frac{1}{t}\left[\sum_{i=1}^{t} \boldsymbol{p}_i\right] - \frac{1}{t}\left[\sum_{i=0}^{t-1} \boldsymbol{p}_i\right] = \frac{1}{t}(\boldsymbol{p}_t - \boldsymbol{p}_0)$
- $b_t = a_t P - a_t$ satisfies $||b_t||_1 \leq \frac{2}{t} \to 0$

Fundamental Theorem of Markov Chains

- $n \times (n+1)\ matrix\ \boldsymbol{A} = [P - I, \mathbf{1}]$ has rank n
- $n \times n\ matrix\ \boldsymbol{B}$ = last n columns of $\boldsymbol{A}$
- First n columns of $\boldsymbol{A}$ sum to zero $\Rightarrow rank(\boldsymbol{B}) = n$
- $c_{\boldsymbol{t}}$ from $b_{\boldsymbol{t}} = a_{\boldsymbol{t}}P - a_{\boldsymbol{t}}$ by dropping first entry
- $a_{\boldsymbol{t}}B = [c_{\boldsymbol{t}}, 1] \Rightarrow a_{\boldsymbol{t}} = [c_{\boldsymbol{t}}, 1]B^{-1}$
- $b_{\boldsymbol{t}} \to 0 \Rightarrow [c_{\boldsymbol{t}}, 1] \to [\mathbf{0}, 1] \Rightarrow a_{\boldsymbol{t}} \to [\mathbf{0}, 1]B^{-1}$
- Let $[\mathbf{0}, 1]B^{-1} = \boldsymbol{\pi}$.
- Since $a_{\boldsymbol{t}} \to \boldsymbol{\pi}$ vector $\boldsymbol{\pi}$ is a probability distribution
- $a_{\boldsymbol{t}}[P - I] = b_{\boldsymbol{t}} = 0 \Rightarrow \boldsymbol{\pi}[P - I] = 0$

Intro to ML

- Classification problem
  - Instance space X: $\{0,1\}^{\boldsymbol{d}}$ or $\mathbb{R}^{\boldsymbol{d}}$ (feature vectors)
  - Classification: come up with a mapping $X \rightarrow \{0,1\}$
- Formalization:
  - Assume there is a probability distribution D over X
  - $\boldsymbol{c}^*$= “target concept” (set $\boldsymbol{c}^* \subseteq X$ of positive instances)
  - Given labeled i.i.d. samples from D produce $\boldsymbol{h} \subseteq X$
  - **Goal:** have $\boldsymbol{h}$ agree with $\boldsymbol{c}^*$ over distribution D
  - Minimize: $err_D(\boldsymbol{h}) = \Pr_D[\boldsymbol{h} \, \Delta \, \boldsymbol{c}^*]$
  - $err_D(\boldsymbol{h})$ = “true” or “generalization” error

Intro to ML

- Training error
  - S = labeled sampled (pairs $(x, l), x \in X, l \in \{0,1\}$)
  - Training error: $err_S(\boldsymbol{h}) = \frac{|S \cap (\boldsymbol{h} \, \Delta \, \boldsymbol{c}^*)|}{|S|}$
- "Overfitting": low training error, high true error
- Hypothesis classes:
  - H: collection of subsets of X called hypotheses
    - If $X = \mathbb{R}$ could be all intervals $\{[a, b], a \leq b\}$
    - If $X = \mathbb{R}^d$ could be linear separators:
      $\left\{\{\boldsymbol{x} \in \mathbb{R}^d \middle| \boldsymbol{w} \cdot \boldsymbol{x} \geq w_0\} | \boldsymbol{w} \in \mathbb{R}^d, w_0 \in \mathbb{R}\right\}$
- If S is large enough (compared to some property of H) then overfitting doesn't occur

Overfitting and Uniform Convergence

- **PAC learning (agnostic)**: For $\epsilon, \delta > 0$ if

$$|S| \geq 1/2\epsilon^2(\ln|H| + \ln 2/\delta)$$

then with probability $1 - \delta$:

$$\forall \boldsymbol{h} \in \mathrm{H}: |err_S(\boldsymbol{h}) - err_D(\boldsymbol{h})| \leq \epsilon$$

- $x_j =$ r.v. (=1 if $\boldsymbol{h}$ has error on j-th sample in S)
- $\mathbb{E}[x_j] = err_D(\boldsymbol{h})$ and $err_S(\boldsymbol{h}) = \frac{1}{|S|}\sum_{j=1}^{|S|} x_j$
- Chernoff bound:

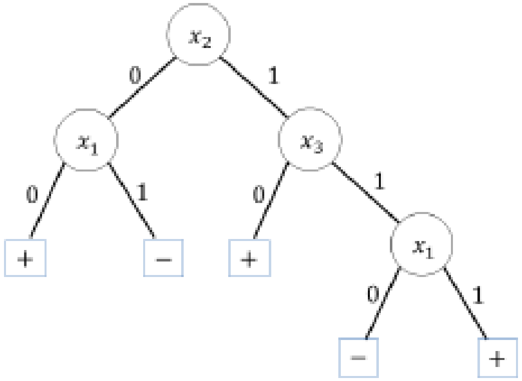
$$\Pr[|err_S(\boldsymbol{h}) - err_D(\boldsymbol{h})| > \epsilon] \leq 2e^{-2|S|\epsilon^2}$$

- Union bound:

$$\Pr[\exists \boldsymbol{h} \in H: |err_S(\boldsymbol{h}) - err_D(\boldsymbol{h})| > \epsilon] \leq 2|H|e^{-2|S|\epsilon^2} \leq \delta$$

Examples

- Learning disjunctions
  - $X = \{0,1\}^{d}$ target concept is OR: $\vee_{i \in T} x_i$
  - $|H| = 2^{d}$ so $|S| = 1/2\epsilon^2(d \ln 2 + \ln 2/\delta)$
- Occam's razor:
  - Target concept can be described by $\leq b$ bits
  - $|H| = 2^{b}$ so $|S| = 1/2\epsilon^2(b \ln 2 + \ln 2/\delta)$
- Learning decision trees
  - $X = \{0,1\}^{d}$
  - $|H|$ = trees with k nodes
  - Described with $b = O(k \log d)$ bits



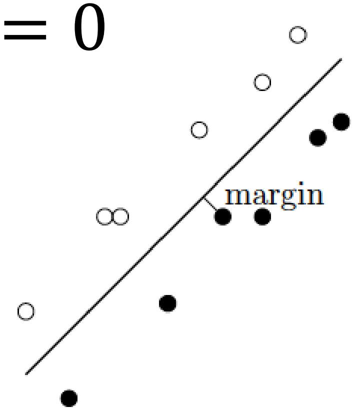
Online Learning + Perceptron Algorithm

- For $t = 1, 2, \ldots,$
  - Algorithm given $x_t \in X$ and asked to predict l_t
  - Algorithm is told $\boldsymbol{c}^*(x_t)$ and charged if $\boldsymbol{c}^*(x_t) \neq l_t$
- Linear separator given by $\boldsymbol{w}^* \in \mathbb{R}^d$

$$\{\boldsymbol{x} \in \mathbb{R}^d \,|\, \boldsymbol{x}^T \boldsymbol{w}^* \geq 1\} = \text{positive examples}$$

$$\{\boldsymbol{x} \in \mathbb{R}^d \,|\, \boldsymbol{x}^T \boldsymbol{w}^* \leq -1\} = \text{negative examples}$$

- $\boldsymbol{x}^T \boldsymbol{w}^* / ||\boldsymbol{w}^*||_2$ = distance to hyperplane $\boldsymbol{x}^T \boldsymbol{w}^* = 0$
- $\gamma = 1/||\boldsymbol{w}^*||_2$ = "margin" of the separator



Perceptron Algorithm

- Set $\boldsymbol{w} = 0$ then for $t = 1,2, \ldots$:
  - Given example x_t predict $\mathrm{sgn}(\boldsymbol{x}_t^T \boldsymbol{w})$
  - If mistake was made then update:
    - If x_t was positive: $\boldsymbol{w} \leftarrow \boldsymbol{w} + \boldsymbol{x_t}$
    - If x_t was negative: $\boldsymbol{w} \leftarrow \boldsymbol{w} - \boldsymbol{x_t}$
- **Thm.** Perceptron makes $\leq R^2 \left|\left|\boldsymbol{w}^*\right|\right|_2^2$ mistakes where $R = \max_t \left|\left|\boldsymbol{x}_t\right|\right|$.
- **Proof:** invariants $\boldsymbol{w^T} \boldsymbol{w}^*$ and $\left|\left|\boldsymbol{w}\right|\right|^{\boldsymbol{2}}$
- For each mistake $\boldsymbol{w^T} \boldsymbol{w}^* \rightarrow \boldsymbol{w^T} \boldsymbol{w}^* + 1$
  - On positive: $(\boldsymbol{w} + \boldsymbol{x}_t)^T \boldsymbol{w}^* = \boldsymbol{w^T} \boldsymbol{w}^* + \boldsymbol{x}_t^T \boldsymbol{w}^* \geq \boldsymbol{w^T} \boldsymbol{w}^* + 1$
  - On negative: $(\boldsymbol{w} - \boldsymbol{x}_t)^T \boldsymbol{w}^* = \boldsymbol{w^T} \boldsymbol{w}^* - \boldsymbol{x}_t^T \boldsymbol{w}^* \geq \boldsymbol{w^T} \boldsymbol{w}^* + 1$

Perceptron Analysis cont.

- On each mistake $||\boldsymbol{w}||_2^2$ increase by $\leq R^2$
- On positive: $(\boldsymbol{w} + \boldsymbol{x}_t)^T(\boldsymbol{w} + \boldsymbol{x}_t) = ||\boldsymbol{w}||_2^2 + 2\boldsymbol{x}_t^T\boldsymbol{w} + ||\boldsymbol{x}_t||_2^2 \leq ||\boldsymbol{w}||_2^2 + ||\boldsymbol{x}_t||_2^2 = ||\boldsymbol{w}||_2^2 + R^2$
- On negative: $(\boldsymbol{w} - \boldsymbol{x}_t)^T(\boldsymbol{w} - \boldsymbol{x}_t) = ||\boldsymbol{w}||_2^2 - 2\boldsymbol{x}_t^T\boldsymbol{w} + ||\boldsymbol{x}_t||_2^2 \leq ||\boldsymbol{w}||_2^2 + ||\boldsymbol{x}_t||_2^2 = ||\boldsymbol{w}||_2^2 + R^2$
- M mistakes: $\boldsymbol{w}^T\boldsymbol{w}^* \geq M$, $||\boldsymbol{w}||_2^2 \leq MR^2$ or $||\boldsymbol{w}||_2 \leq \sqrt{M}R$
- Since $\frac{\boldsymbol{w}^T\boldsymbol{w}^*}{||\boldsymbol{w}^*||_2} \leq ||\boldsymbol{w}||_2$ we have:

$$\frac{M}{||\boldsymbol{w}^*||_2} \leq \sqrt{M}R \Rightarrow \sqrt{M} \leq R||\boldsymbol{w}^*||_2 \Rightarrow M \leq R^2||\boldsymbol{w}^*||_2^2$$

Perceptron with noisy data

- What if there is no perfect separator?
- Hinge loss of $\boldsymbol{w}^*$:
  - On positive x_t: $\max(0,1 - \boldsymbol{x}_t^T \boldsymbol{w}^*)$
  - On negative x_t: $\max(0,1 + \boldsymbol{x}_t^T \boldsymbol{w}^*)$
- Sample hinge loss $L_{hinge}(\boldsymbol{w}^*, S) =$ sum of hinge losses over all samples in S
- **Thm.** #mistakes of Perceptron is at most:

$$\min_{\boldsymbol{w}^*} \left(R^2 ||\boldsymbol{w}^*\Big|\Big|_2^2 + 2L_{hinge}(\boldsymbol{w}^*, S) \right)$$

Proof of noisy perceptron

- As before we have $||\boldsymbol{w}||_2^2 \leq MR^2$
- On positive: $(\boldsymbol{w} + \boldsymbol{x}_t)^T \boldsymbol{w}^* = \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{w}^* + \boldsymbol{x}_t^T \boldsymbol{w}^* \geq \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{w}^* + 1 - L_{hinge}(\boldsymbol{w}^*, \boldsymbol{x}_t)$
- On negative: $(\boldsymbol{w} + \boldsymbol{x}_t)^T \boldsymbol{w}^* = \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{w}^* - \boldsymbol{x}_t^T \boldsymbol{w}^* \geq \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{w}^* + 1 - L_{hinge}(\boldsymbol{w}^*, \boldsymbol{x}_t)$
- In the end: $\boldsymbol{w}^T \boldsymbol{w}^* \leq M - L_{hinge}(\boldsymbol{w}^*, S)$
- Similar argument as before shows that:

$$M \leq R^2 ||\boldsymbol{w}^*||_2^2 + 2L_{hinge}(\boldsymbol{w}^*, S)$$