Exponential Families
Classification and Novelty Detection

S.V.N. “Vishy” Vishwanathan

SVN.Vishwanathan@nicta.com.au
http://web.rsise.anu.edu.au/~vishy

National ICT Australia
and
Australian National University

Thanks to Alex Smola, Thomas Hofmann and Stéphane Canu

Overview

- Review of Exponential Family
  - Log Partition Function
  - Conditional Densities
  - Missing Variables
- Maximum Likelihood Estimation
- MAP Estimation
  - Gaussian Processes and the Normal Prior
  - Novelty Detection
  - Large Margin Classifiers
- Graphical Models
  - Hammersley Clifford Theorem
  - Conditional Random Fields

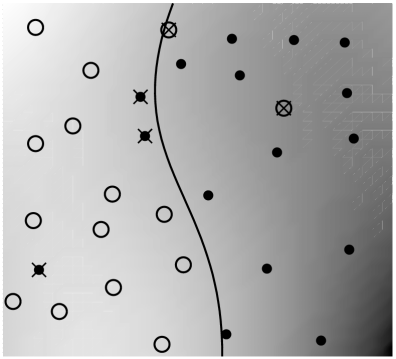
Machine Learning

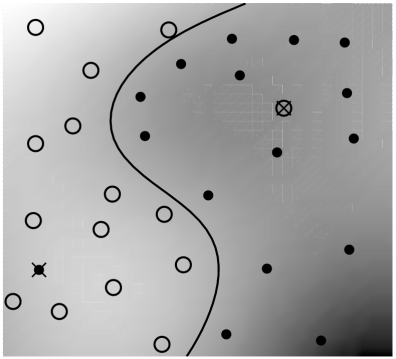
Data:

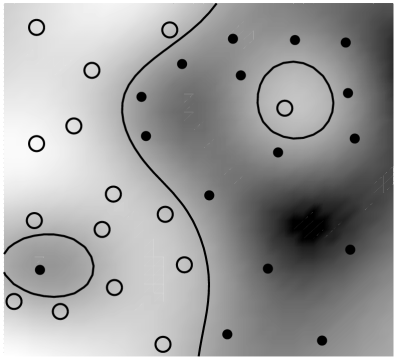
- Pairs of observations $(\mathbf{x}_i, y_i)$
- Underlying distribution $\mathrm{P}(\mathbf{x}, y)$
- Examples (blood status, cancer), (transactions, fraud)

Task:

- Find a function $f(\mathbf{x})$ which predicts y given $\mathbf{x}$
- The function $f(\mathbf{x})$ must *generalize* well







Exponential Family

Basic Equation:

- We will use:

$$p(\mathbf{x};\theta) = p_0(\mathbf{x}) \exp(\langle \phi(\mathbf{x}), \theta \rangle - g(\theta))$$

Why?

- Dense in space of densities (L_∞ sense)
- We can use $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ where $\mathcal{H}$ is a RKHS
- Conditional models and graphical models

Where is the Catch?

- The log-partition function

$$g(\theta) = \log \int_{\mathcal{X}} \exp(\langle \phi(\mathbf{x}), \theta \rangle) \, d\,\mathbf{x}$$

  is difficult to compute

Log-Partition Function

NATIONAL ICT AUSTRALIA

Basic Equation:

- To ensure density integrates to 1

$$g(\theta) = \log \int_{\mathcal{X}} \exp(\langle \phi(\mathbf{x}), \theta \rangle) \, d\,\mathbf{x}$$

- Domain $\mathcal{X}$ might be large (computing integral is painful)

Moment Generating Function:

- Derivatives of $g(\theta)$ generate moments of $\phi(\mathbf{x})$

$$\begin{aligned} \partial_\theta g(\theta) &= \mathbb{E}_{p(\mathbf{x};\theta)} \left[\phi(\mathbf{x}) \right] \\ \partial^2_\theta g(\theta) &= \mathrm{Var}_{p(\mathbf{x};\theta)} \left[\phi(\mathbf{x}) \right] \end{aligned}$$

Corollary:

- The log-partition function is convex
- It is smooth and differentiable (C^∞ function)

Some Examples

Bernoulli Distribution:

- Given by $p(x;\mu) = \mu^x(1-\mu)^{1-x}$
- Exponential family form

$$p(x;\theta) = \exp(\langle x, \theta\rangle - \log(1+\exp(\theta))) \text{ where } \theta = \log\frac{\mu}{1-\mu}$$

Laplace Distribution:

- Model the decay of atoms by $p(x;\theta) = \theta\exp(\langle -x, \theta\rangle)$
- In exponential family form

$$p(x;\theta) = \exp(\langle -x, \theta\rangle - (-\log\theta))$$

Poisson Distribution:

- In exponential family form

$$p(x) = \exp\left(\langle x, \theta\rangle - \log\Gamma(x+1) - \exp(\theta)\right)$$

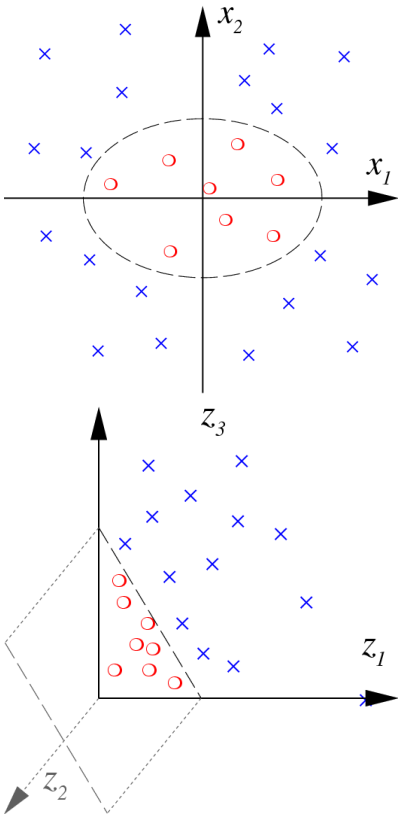
Kernels

Problem: We want to perform non-parametric estimation (i.e. without designing the sufficient statistics)

Idea 1: Map to a higher dimensional feature space via $\Phi : x \rightarrow \Phi(x)$ and solve the problem there Replace every $\langle x, x' \rangle$ by $\langle \Phi(x), \Phi(x') \rangle$

Idea 2: Instead of computing $\Phi(x)$ explicitly use a **kernel function** $k(x, x') := \langle \Phi(x), \Phi(x') \rangle$

- A large class of functions are admissible as kernels
- Non-vectorial data can be handled if we can compute meaningful $k(x, x')$



Unconditional Models

Likelihood of data:

- Data $\mathbf{X} := \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ generated IID
- Likelihood is given by

$$p(\mathbf{X}; \theta) = \prod_{i=1}^{m} p(\mathbf{x}_i; \theta) = \exp\left(\sum_{i=1}^{m} \langle \phi(\mathbf{x}_i), \theta \rangle - m g(\theta)\right)$$

Maximum Likelihood:

- We want to minimize the negative log-likelihood

$$\underset{\theta}{\text{minimize}} \quad g(\theta) - \left\langle \frac{1}{m} \sum_{i=1}^{m} \phi(\mathbf{x}_i), \theta \right\rangle$$

$$\implies \quad \mathbf{E}[\phi(\mathbf{x})] = \frac{1}{m} \sum_{i=1}^{m} \phi(\mathbf{x}_i) =: \mu$$

An Example

Estimate the decay constant of an atom:

- We model decay using Laplace distribution
- Using exponential family notation

$$p(x;\theta) = \exp(\langle -x, \theta \rangle - (-\log \theta))$$

Computing μ:

- Observe that $\phi(x) = -x$
- All we need to do is average over observed decay times!

Solving for Maximum Likelihood:

- The maximum likelihood condition yields

$$\mu = \partial_\theta g(\theta) = \partial_\theta(-\log \theta) = -\frac{1}{\theta}$$

- This leads to $\theta = -\frac{1}{\mu}$

Conditional Models

Why Conditional Models?

- We are given $(\mathbf{x}_i, y_i)$ pairs
- Given a new data point we want to predict its label y
- We don't want to waste modeling effort on $\mathbf{x}$

Conditional Exponential Family:

- Assume $p(y|\,\mathbf{x};\theta)$ is a member of the exponential family

$$p(y|\,\mathbf{x};\theta) = \exp(\langle\phi(\mathbf{x},y),\theta\rangle - g(\theta|\,\mathbf{x}))$$

$$g(\theta|\,\mathbf{x}) = \log\int_y \exp(\langle\phi(\mathbf{x},\bar{\mathbf{y}}),\theta\rangle\; d\,\bar{\mathbf{y}}$$

The Task Ahead:

- Estimate parameter θ and compute $g(\theta|\,\mathbf{x}))$

Handling Missing Variables

Partially Observed Data:

- For partially observed data

$$p(\mathbf{x}^u, y | \mathbf{x}^o; \theta) = \exp\left(\langle \phi(\mathbf{x}^o, \mathbf{x}^u, y), \theta \rangle - g(\theta | \mathbf{x}^o)\right).$$

The Density:

- Integrate out unobserved part

$$\begin{aligned} p(y | \mathbf{x}^o; \theta) &= \int_{\mathcal{X}^u} \exp\left(\langle \phi(\mathbf{x}^o, \mathbf{x}^u, y), \theta \rangle - g(\theta | \mathbf{x}^o)\right) d\mathbf{x}^u \\ &= \exp(g(\theta | \mathbf{x}^o, y) - g(\theta | \mathbf{x}^o)) \end{aligned}$$

Overview

- Review of Exponential Family
  - Log Partition Function
  - Conditional Densities
  - Missing Variables
- Maximum Likelihood Estimation
- MAP Estimation
  - Gaussian Processes and the Normal Prior
  - Novelty Detection
  - Large Margin Classifiers
- Graphical Models
  - Hammersley Clifford Theorem
  - Conditional Random Fields

Max-Likelihood Estimation

Observe Data:

- Data $\mathbf{X} = \{(\mathbf{x}_i, y_i)\}$ drawn from distribution $p(\mathbf{x}, y|\theta)$

Compute Likelihood:

- Since data is assumed IID

$$\log p(y|\,\mathbf{X};\theta) = \sum_{i=1}^{m} \log p(y_i|\,\mathbf{x}_i;\theta)$$

Maximize it:

- Take the negative log and minimize, which leads to

$$\mathbb{E}_{p(y|\,\mathbf{x};\theta)}[\phi(\mathbf{x}, y)] = \frac{1}{m}\sum_{i=1}^{m} \phi(\mathbf{x}_i, y_i)$$

- This can be solved analytically or by Newton's method.

Overview

- Review of Exponential Family
  - Log Partition Function
  - Conditional Densities
  - Missing Variables
- Maximum Likelihood Estimation
- MAP Estimation
  - Gaussian Processes and the Normal Prior
  - Novelty Detection
  - Large Margin Classifiers
- Graphical Models
  - Hammersley Clifford Theorem
  - Conditional Random Fields

MAP Estimation

A True Bayesian:

- We assume that θ is a random variable
- Also assume a prior (belief) over θ

The Normal Prior:

- We assume $\theta \sim \mathcal{N}(0, \sigma^2)$
- The posterior (Bayes rule)

$$p(\theta|\,\mathbf{X}, y) \propto \exp\left(\sum_{i=1}^{m} \langle \phi(\mathbf{x}_i, y), \theta \rangle - g(\theta|\,\mathbf{x}) - \frac{1}{2\sigma^2}\|\theta\|^2\right)$$

The Solution:

- By setting $\partial_\theta - \log p(\theta|\,\mathbf{X}, y) = 0$ we get

$$\mathbb{E}_{p(y|\,\mathbf{x};\theta)}[\phi(\mathbf{x}, y)] = \frac{1}{m}\sum \phi(\mathbf{x}_i, y_i) - \frac{\theta}{m\sigma^2}$$

Gaussian Processes

NATIONAL ICT AUSTRALIA

Key Idea:

- Let $t : \mathcal{X} \to \mathbb{R}$ be a stochastic process
- Fix any $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$
- For a GP $\{t(\mathbf{x}_1), \ldots, t(\mathbf{x}_m)\}$ are jointly normal

Parameters of a GP:

- Mean
$$\mu(\mathbf{x}) := \mathbb{E}[t(\mathbf{x})]$$
- Covariance function (kernel)
$$k(\mathbf{x}, \mathbf{x}') := \mathrm{Cov}(t(\mathbf{x}), t(\mathbf{x}'))$$

Simplifying Assumption:

- Mean $\mu(\mathbf{x}) = 0$
- We know the form of $k(\mathbf{x}, \mathbf{x}')$

Our Model and GP

Key Idea:

- Let $\theta \sim \mathcal{N}(0, \sigma^2)$
- Then $\log p(y|\,\mathbf{x}; \theta) + g(\theta|\,\mathbf{x})$ is a GP

Why?:

- Observe that $\log p(y|\,\mathbf{x}; \theta) + g(\theta|\,\mathbf{x}) = \langle \phi(\mathbf{x}, y), \theta \rangle$
- Hence it is normally distributed
- The mean $\mathbb{E}_\theta[\langle \phi(\mathbf{x}, y), \theta \rangle] = 0$
- The covariance function

$$k((\mathbf{x}, y), (\mathbf{x}', y)) = \sigma^2 \langle \phi(\mathbf{x}, y), \phi(\mathbf{x}', y') \rangle$$

Observations:

- Kernel can depend on both $\mathbf{x}$ and y
- Extensions to multi-class problems possible
- If y has structure we can exploit it

Representer Theorem

Optimization Problem:

- The MAP estimate solves

$$\operatorname*{argmin}_{\theta} \frac{1}{2\sigma^2}||\theta||^2 + \sum_{i=1}^{m} g(\theta|\,\mathbf{x}_i) - \langle \phi(\mathbf{x}_i, y_i), \theta \rangle$$

Representer Theorem:

- By the representer theorem

$$\theta = \sum_{i=1}^{m} \sum_{y \in \mathcal{Y}} \alpha_{iy} \phi(\mathbf{x}_i, y_i)$$

Observations:

- If $|\,\mathcal{Y}\,|$ is large we are in trouble
- For binary classification we will use $\phi(\mathbf{x}, y) = y\phi(\mathbf{x})$

GP and Missing Variables

Partially Observed Data:

- Plug in density into MAP estimate

$$\text{minimize} \sum_{i=1}^{m} [g(\theta|\, \mathbf{x}_i^o) - g(\theta|\, \mathbf{x}_i^o, y_i)] + \frac{1}{2\sigma^2}\|\theta\|^2$$

Can of Worms?

- Optimization problem is no longer convex
- We will come back to this . . .

Conjugate Priors

Problems with Normal Prior:

- Posterior looks different from prior
- Many MLE algorithms may not work

Idea:

- What if the prior looked like additional data

$$p(\theta|\,\mathbf{X}) \sim p(\mathbf{X}\,|\theta)$$

- For exponential families set

$$p(\theta|a) \propto \exp(\langle m_0 a, \theta\rangle - m_0 g(\theta))$$

The Posterior:

- Now looks like

$$p(\theta|\,\mathbf{X}) \propto \exp\left((m+m_0)\left(\left\langle \frac{m\mu + m_0 a}{m+m_0}, \theta \right\rangle - g(\theta)\right)\right)$$

Optimization Problems

Maximum Likelihood:

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} g(\theta) - \langle \phi(\mathbf{x}_i), \theta \rangle \Longrightarrow \partial_\theta g(\theta) = \frac{1}{m} \sum_{i=1}^{m} \phi(\mathbf{x}_i)$$

Normal Prior:

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} g(\theta) - \langle \phi(\mathbf{x}_i), \theta \rangle + \frac{1}{2\sigma^2} \|\theta\|^2$$

Conjugate Prior:

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} g(\theta) - \langle \phi(\mathbf{x}_i), \theta \rangle + m_0 g(\theta) - m_0 \langle a, \theta \rangle$$

equivalently solve $\partial_\theta g(\theta) = \frac{1}{m + m_0} \sum_{i=1}^{m} \phi(\mathbf{x}_i) + \frac{m_0}{m + m_0} a$

Novelty Detection

Key Idea:

- We estimate $p(\mathbf{x}\,|\theta)$ based on $\{\mathbf{x}_i\}$
- All $\mathbf{x}_i$ with $p(\mathbf{x}_i\,|\theta) < p_0$ are novel

Tightening the Belt:

- Don't waste modeling effort on high density regions
- Only shape of $p(\mathbf{x}\,|\theta)$ is important

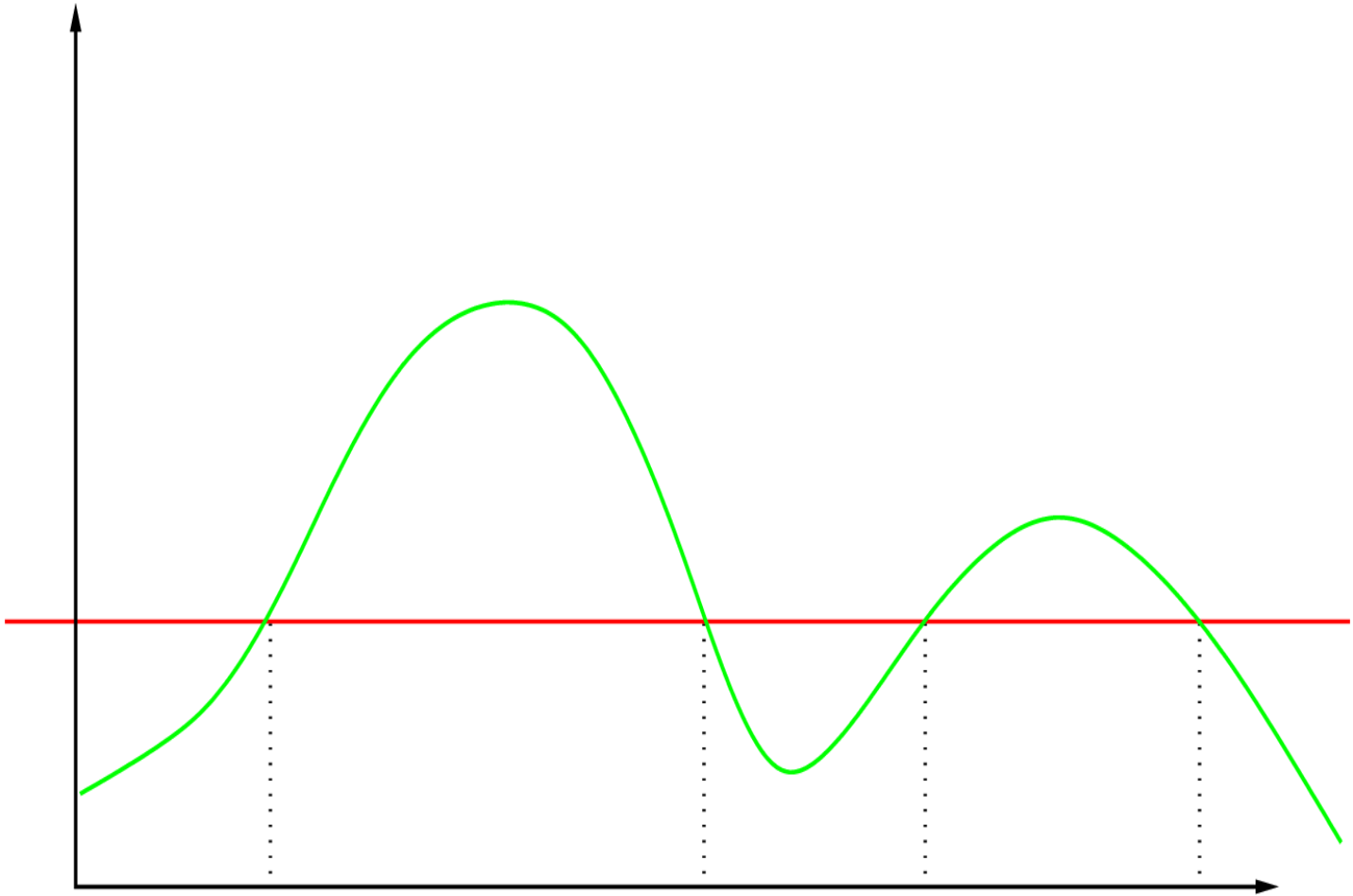
The Solution:

- Estimate

$$\min\left(\frac{p(\mathbf{x}_i\,|\theta)}{p_0}, 1\right)$$

- We use $p_0 = \exp(\rho - g(\theta))$
- Helps get rid of pesky $g(\theta)$ term

Novelty Detection



Optimization Problem

Exponential Family:

- Using the iid assumption our objective function is

$$\operatorname*{argmax}_{\theta} \prod_{i=1}^{m} \min\left(\frac{p(\mathbf{x}_i\,|\theta)}{p_0}, 1\right) p(\theta)$$

The Final Form:

- If we assume a normal prior and use log likelihoods

$$\operatorname*{argmin}_{\theta} \sum_{i=1}^{m} \max(\rho - \langle \phi(\mathbf{x}_i), \theta \rangle, 0) + \frac{1}{2\sigma^2}||\theta||^2$$

- Exactly the problem solved by the Single class SVM!

The ν-Trick:

- Assume $p(\rho) \propto \exp(\nu m \rho)$

Odds Ratio

Basic Idea:

- In OCR classification 0 and 8 are frequently confused
- Digits like 0 and 1 are generally well classified
- Worst confused class is a measure of margin

The Solution:

- If we consider the ratio

$$R(\mathbf{x}, y, \theta) = \log \frac{p(y|\,\mathbf{x}, \theta)}{\max_{y \neq y'} p(y'|\,\mathbf{x}, \theta)}$$

- Measure of confusion with the next best class
- We can interpret this as the margin

The Consequences:

- SVM like large margin classifiers are special cases
- Extensions to multi-class setting natural

Multiclass SVM

A Problem:

- Three classes $\{1, 2, 3\}$
- Two densities $\{0.3, 0.3, 0.4\}$ and $\{0.1, 0.5, 0.4\}$
- Class 3 has same likelihood in both cases
- It is misclassified in the second case!

Odds Ratio Loss:

- The log odds ratio behaves like a margin

$$\begin{aligned} R(\mathbf{x}, y, \theta) &= \log \frac{p(y|\,\mathbf{x}, \theta)}{\max_{y \neq y'} p(y'|\,\mathbf{x}, \theta)} \\ &= \langle \phi(\mathbf{x}, y), \theta \rangle - \max_{y \neq y'} \langle \phi(\mathbf{x}, y'), \theta \rangle \end{aligned}$$

- Generalized hinge loss

$$c(\mathbf{x}, y, \theta) := \max(1 - R(\mathbf{x}, y, \theta), 0)$$

Multiclass SVM - II

Optimization Problem:

- We solve

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}||\theta||^2 \\ &\text{s.t. } R(\mathbf{x}_i, y_i, \theta) \geq 1 \end{aligned}$$

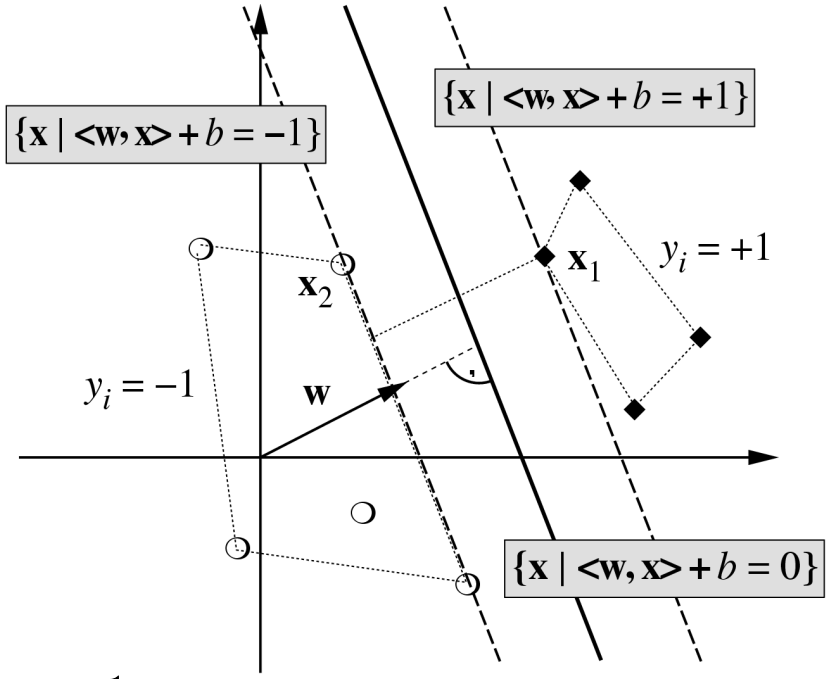
Binary SVM:

- Set $\phi(\mathbf{x}, y) = \frac{y}{2}\phi(\mathbf{x})$
- Then $R(\mathbf{x}, y, \theta) = y \langle \phi(\mathbf{x}), \theta \rangle$
- We now solve

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}||\theta||^2 \\ &\text{s.t. } y_i \langle \phi(\mathbf{x}_i), \theta \rangle \geq 1 \end{aligned}$$

- This is exactly the hard margin binary SVM!

Optimal Separating Hyperplane



Note:

$$\langle \mathbf{w}, \mathbf{x}_1 \rangle + b = +1$$
$$\langle \mathbf{w}, \mathbf{x}_2 \rangle + b = -1$$
$$\Rightarrow \quad \langle \mathbf{w}, (\mathbf{x}_1 - \mathbf{x}_2) \rangle = 2$$
$$\Rightarrow \left\langle \frac{\mathbf{w}}{\|\mathbf{w}\|}, (\mathbf{x}_1 - \mathbf{x}_2) \right\rangle = \frac{2}{\|\mathbf{w}\|}$$

Minimize $\frac{1}{2}\|\theta\|^2$ subject to $y_i \langle \theta, \mathbf{x}_i \rangle \geq 1$ for all i

Odds Ratio and Missing Vars

NATIONAL ICT AUSTRALIA

Log Odds Ratio:

- From definition of $p(y|\,\mathbf{x}^o;\theta)$ and $R(\mathbf{x},y,\theta)$:

$$R(\mathbf{x},y,\theta) = g(\theta|\,\mathbf{x}^o,y) - \max_{y'\neq y} g(\theta|\,\mathbf{x}^o,y')$$

- Expensive to compute!

The Optimization Problem:

- We now have to solve

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}||\theta||^2 \\ &\text{s.t. } g(\theta|\,\mathbf{x}^o,y) - \max_{y'\neq y} g(\theta|\,\mathbf{x}^o,y') \geq 1 \end{aligned}$$

The Challenge:

- Constraints are not convex
- Clever sampling techniques required

Soft Margin SVM's

Slack Variables:

- Data might not be linearly separable in feature space
- To avoid over-fitting ignore noisy points
- We modify the optimization problem

$$\min \tfrac{1}{2}||\theta||^2 + C\sum_i \xi_i$$
$$\text{s.t.} R(\mathbf{x}_i, y_i, \theta) \geq 1 - \xi_i \quad \xi_i \geq 0$$

Upper Bound on Error:

- If we define

$$\xi_i(\theta) = \max\{0, 1 - R(\mathbf{x}_i, y_i, \theta)\}$$

then

$$\frac{1}{m}\sum_{i=1}^{m} \xi_i(\theta) \geq \frac{1}{m}\sum_{i=1}^{m} \delta(y_i, \text{sign}(\log R(\mathbf{x}_i, y_i, \theta)))$$

Another Formulation

Slack Variables:

- We include a slack term for every linear constraint
- The optimization problem becomes

$$\min \tfrac{1}{2}||\theta||^2 + C\sum_i \sum_{y \neq y_i} \xi_{iy}$$
$$\text{s.t.} \langle \phi(\mathbf{x}_i, y_i) - \phi(\mathbf{x}_i, y), \theta \rangle \geq 1 - \xi_{iy} \quad \xi_{iy} \geq 0$$

Upper Bound on Ranking Error:

- Now we can write a bound

$$\frac{1}{m}\sum_{i=1}^{m} \xi_{iy}(\theta) \geq \frac{1}{m}\sum_{i=1}^{m} |\{y \neq y_i : \langle \phi(\mathbf{x}_i, y), \theta \rangle \geq \langle \phi(\mathbf{x}_i, y_i), \theta \rangle\}|$$

Comments:

- More constraints $\Longrightarrow$ harder problem to solve
- Solution might not be sparse!

A Laundry List

NATIONAL ICT AUSTRALIA

Gaussian Process:

$$\text{minimize} \sum_{i=1}^{m} [g(\theta|\, \mathbf{x}_i^o) - g(\theta|\, \mathbf{x}_i^o, y_i)] + \frac{1}{2\sigma^2}\|\theta\|^2.$$

SVM with Slack:

$$\text{minimize} \frac{1}{2\sigma^2}\|\theta\|^2 + \sum_{i=1}^{m} \xi_i$$

$$\text{s.t. } 1 - \xi_i - g(\theta|\, \mathbf{x}_i^o, y_i) + \max_{\tilde{y} \neq y_i} g(\theta|\, \mathbf{x}_i^o, \tilde{y}) \leq 0, \quad \text{and} \quad -\xi_i \leq 0$$

Novelty Detection:

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2\sigma^2}\|\theta\|^2 + \textstyle\sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & \rho - \xi_i - g(\theta|\, \mathbf{x}^o) \leq 0 \text{ and } -\xi_i \leq 0 \end{aligned}$$

Constrained CCCP

NATIONAL ICT AUSTRALIA

The Problem:

- Suppose we want to solve

$$\begin{aligned} \text{minimize} \quad & f_0(\mathbf{x}) - g_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) - g_i(\mathbf{x}) \leq c_i \end{aligned}$$

- Both f_i and g_i are convex for all i

The Basic Idea:

- Replace g_i by a linear approximation
- Solve the new convex problem
- Repeat until convergence

Challenges:

- What is a good linear approximation?
- We use first order Taylor approximation
- Rates of convergence?

Universal Density Estimators

NATIONAL ICT AUSTRALIA

Setting:

- Let $\mathcal{X}$ be a measurable set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a kernel
- Let $f(\cdot) = \langle \phi(\cdot), \theta \rangle_{\mathcal{H}}$ and $f(\mathbf{x}) = \langle f(\cdot), k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}$
- The set of continuous and bounded densities on $\mathcal{X}$ be $\mathcal{P}$
- Furthermore let $\mathcal{H}$ be dense in $C^0(\mathcal{X})$

Universal Density Estimators:

- The densities $p_f(\mathbf{x}) := \exp(f(\mathbf{x}) - g_f(\theta))$ are dense in $\mathcal{P}$

Proof Sketch:

- Find a $f(\mathbf{x})$ *close* to given $\bar{\mathbf{p}}(\mathbf{x})$
- Show that $\int_{\mathcal{X}} \exp(f(\mathbf{x}))\, d\,\mathbf{x}$ is bounded
- It follows that $|\log p_f(\mathbf{x}) - \log \bar{\mathbf{p}}(\mathbf{x})|$ is small
- Hence conclude that $|p_f(\mathbf{x}) - \bar{\mathbf{p}}(\mathbf{x})|$ is small

Graphical Models

Basic Idea:

- $\mathbf{x}, \mathbf{x}'$ are conditionally independent given c, if

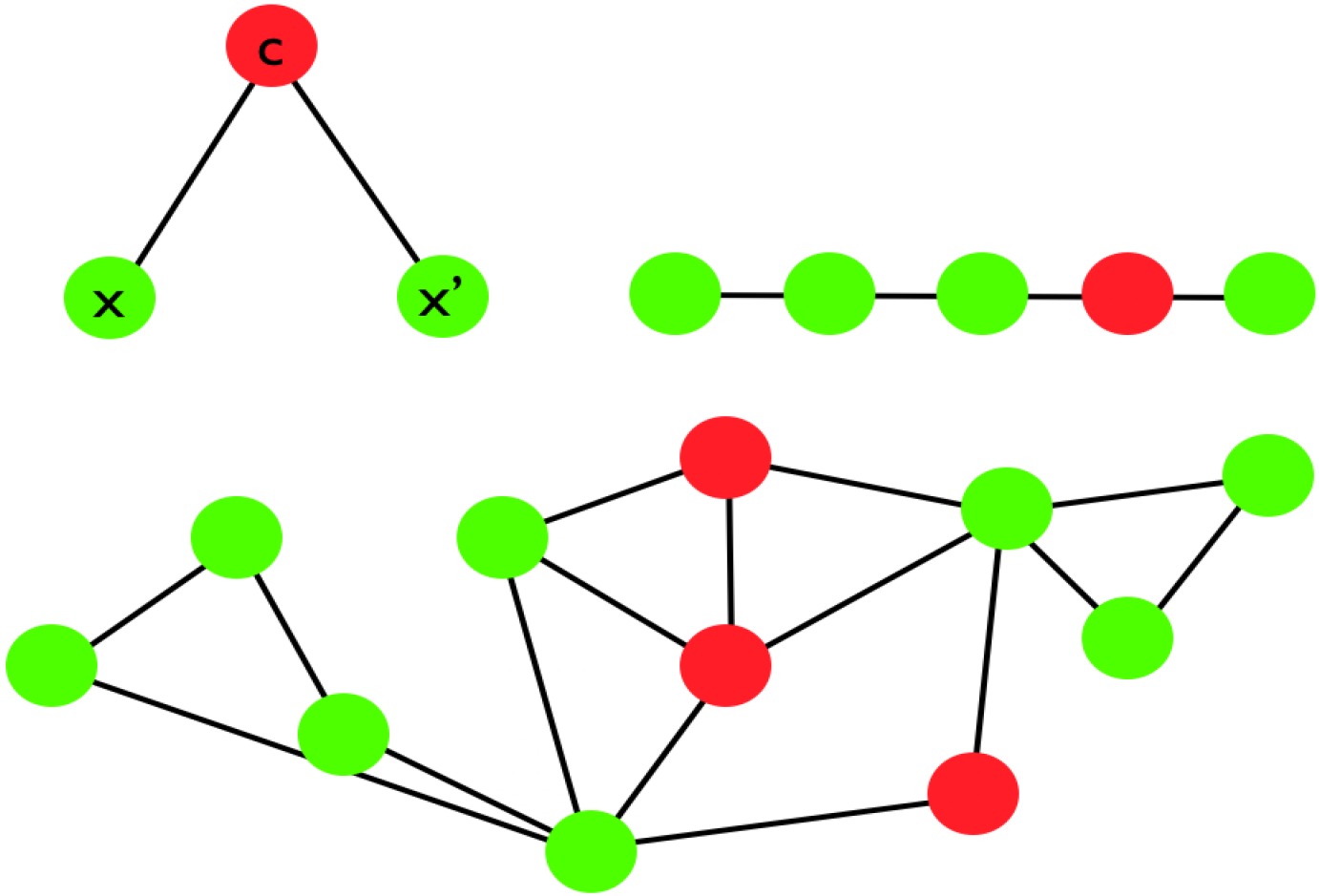
$$p(\mathbf{x}, \mathbf{x}' \,|c) = p(\mathbf{x} \,|c) p(\mathbf{x}' \,|c)$$

- This information might be known in advance
- Can simplify distributions significantly

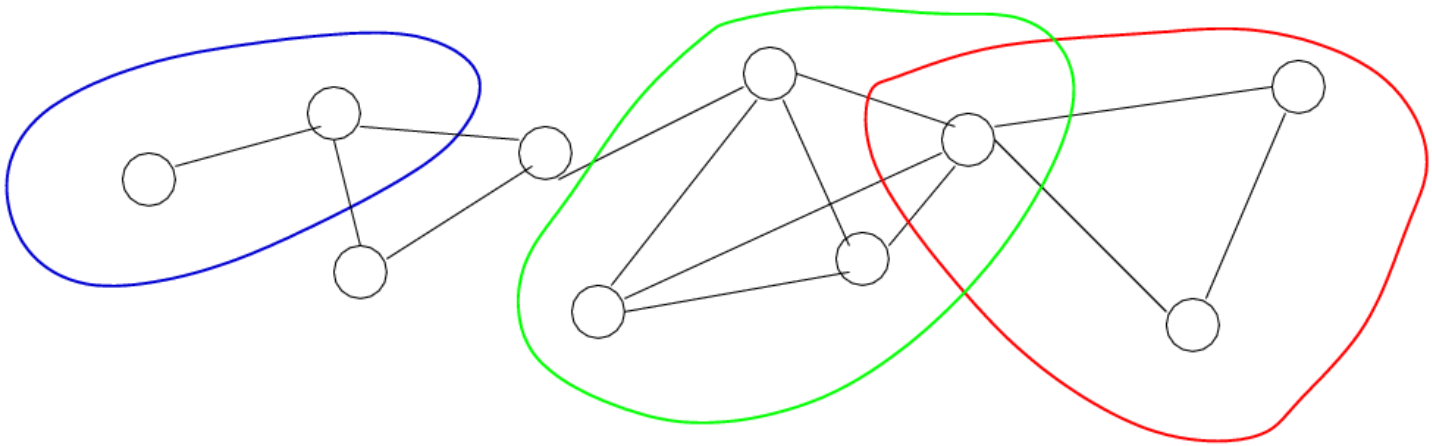
Markov Network:

- A graph $G(V, E)$ with vertices V and edges E
- Each node of G represents a random variable
- Each subgraph of G is a subset of random variables
- Subsets $\mathbf{x}_S, \mathbf{x}_{S'}$ are conditionally independent given x_C if removing the vertices C from G decomposes the graph into disjoint subsets containing S, S'.

Conditional Independence



Maximal Cliques



Definition:

- Maximally connected subgraph of a graph

Why are they Useful?

- Allow us to specify dependencies naturally
- Graph algorithms can be used for inference
- Hammersley Clifford Theorem

Hammersley Clifford Theorem

Theorem:

- If G encodes conditional independence assumptions

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c)\right)$$

Exponential Family:

- Can apply decomposition to the exponential family
- The vector $\phi(\mathbf{x})$ decomposes as

$$\phi(\mathbf{x}) = (\ldots, \phi_c(\mathbf{x}_c), \ldots)$$

Kernel Function:

- The kernel now looks like

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle = \sum_c \langle \phi_c(\mathbf{x}_c), \phi_c(\mathbf{x}'_c) \rangle = \sum_c k(\mathbf{x}, \mathbf{x}')$$

Proof

Step 1: Obtain linear functional

- Combine exponential family with CH theorem:

$$\langle \Phi(\mathbf{x}), \theta \rangle = \sum_{c \in C} \psi_c(\mathbf{x}_c) - \log Z + g(\theta) \text{ for all } \mathbf{x}, \theta$$

Step 2: Orthonormal basis in θ

- Swallow Z and $g(\theta)$
- Expand using an orthonormal basis

$$\langle \Phi(\mathbf{x}), \mathbf{e}_i \rangle = \sum_{c \in C} \eta_c^i(\mathbf{x}_c) \text{ for some } \eta_c^i(\mathbf{x}_c)$$

Proof contd . . .

Step 3: Reconstruct sufficient statistics

- Since

$$\Phi_c(\mathbf{x}_c) := (\eta_c^1(\mathbf{x}_c), \eta_c^2(\mathbf{x}_c), \ldots)$$

we can write

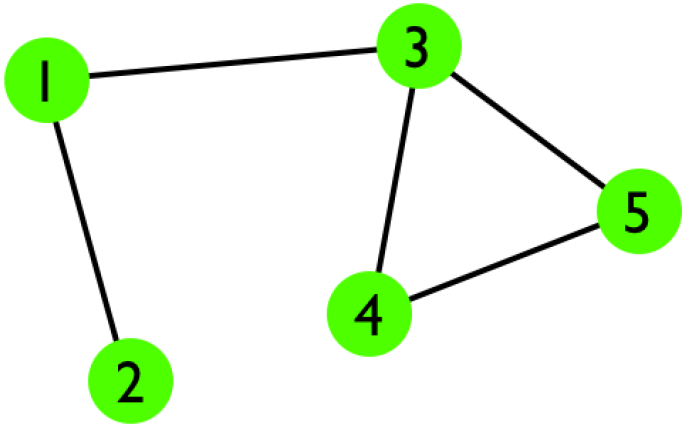
$$\langle \Phi(\mathbf{x}), \theta \rangle = \sum_{c \in C} \sum_i \theta_i \Phi_c^i(\mathbf{x}_c)$$

Example: Normal Distribution

Density:

$$p(\mathbf{x}\,|\theta) = \exp\left(\sum_{i=1}^{n} \mathbf{x}_i\,\theta_{1i} + \sum_{i,j=1}^{n} \mathbf{x}_i\,\mathbf{x}_j\,\theta_{2ij} - g(\theta)\right)$$

Here $\theta_2 = \Sigma^{-1}$, is the inverse covariance matrix. We have that $(\Sigma^{-1})_{[ij]} \neq 0$ only if (i, j) share an edge.



	1	2	3	4	5
1	■	■			
2	■	■	■		
3		■	■	■	■
4			■	■	■
5			■	■	■

Example: Gaussian Distribution

NATIONAL ICT AUSTRALIA

Sufficient Statistics:

- For Gaussian distribution $\phi(\mathbf{x}) = (\mathbf{x}, \mathbf{x}\,\mathbf{x}^\top)$

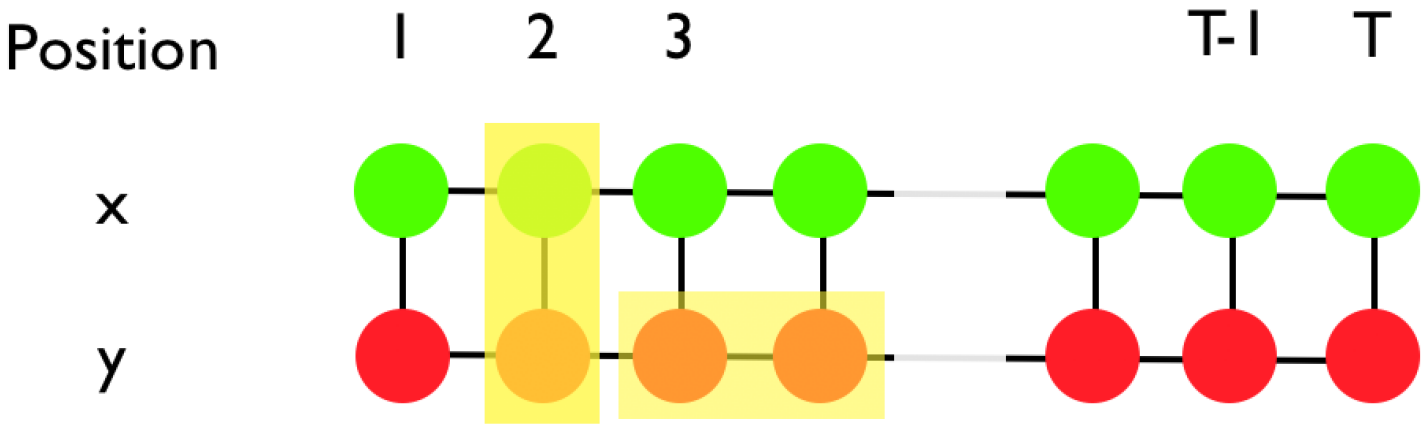
Clifford Hammersley Theorem:

- $\phi(\mathbf{x})$ must decompose into subsets on max cliques
- The linear term trivial to decompose
- An edge in the graph $G(V, E)$ implies coupling
- If $\mathbf{x}_i$ is connected by edge to $\mathbf{x}_j$ then this implies coupling

Inverse Covariance Matrix:

- Inverse covariance matrix corresponds to $\mathbf{x}\,\mathbf{x}^\top$
- Its sparsity mirrors $G(V, E)$
- Sparse inverse kernel matrix corresponds to a graphical model!

Conditional Random Fields



Cliques and Density:

- Cliques are $(\mathbf{x}_t, y_t)$, $(\mathbf{x}_t, \mathbf{x}_{t+1})$, and (y_t, y_{t+1})
- Drop cliques in $(\mathbf{x}_t, \mathbf{x}_{t+1})$: no effect on $p(y|\,\mathbf{x}, \theta)$

$$p(y|\,\mathbf{x}, \theta) = \exp\Big(\sum_t \langle \phi_{xy}(\mathbf{x}_t, y_t), \theta_{xy,t} \rangle + \langle \phi_{yy}(y_t, y_{t+1}), \theta_{yy,t} \rangle + \langle \phi_{xx}(\mathbf{x}_t, \mathbf{x}_{t+1}), \theta_{xx,t} \rangle - g(\theta|\,\mathbf{x}) \Big)$$

Computational Issues

Assumption:

- Assume stationarity of the model
- The various θ_c are time invariant

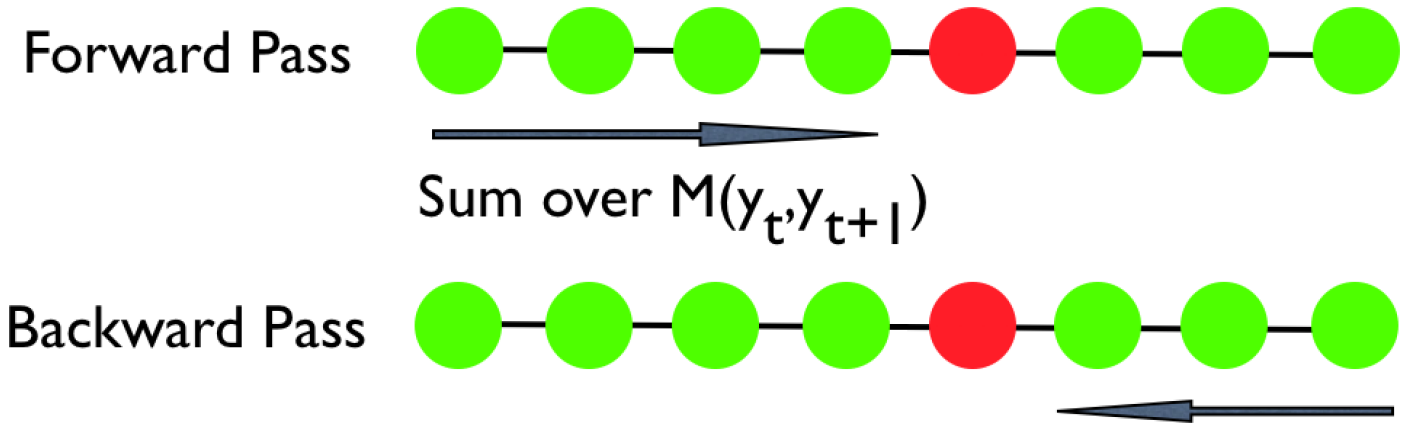
Dynamic Programming:

- Compute $g(\theta|\,\mathbf{x})$ via dynamic programming

$$
\begin{aligned}
& g(\theta|\,\mathbf{x}) \\
&= \log \sum_{y_1,\ldots,y_T} \prod_{t=1}^{T} \underbrace{\exp\left(\langle \phi_{xy}(\mathbf{x}_t, y_t), \theta_{xy} \rangle + \langle \phi_{yy}(y_t, y_{t+1}), \theta_{yy} \rangle\right)}_{M_t(y_t, y_{t+1})} \\
&= \log \sum_{y_1} \sum_{y_2} M_1(y_1, y_2) \sum_{y_3} M_2(y_2, y_3) \ldots \sum_{y_T} M_T(y_{T-1}, y_T)
\end{aligned}
$$

- Dynamic programming for $p(y_t|\,\mathbf{x}, \theta)$ and $p(y_t, y_{t+1}|\,\mathbf{x}, \theta)$

Forward Backward Algorithm



Key Idea:

- Store sum over all $y_1, \ldots, y_{t-1}$ (forward pass) and over all $y_{t+1}, \ldots, y_T$ as intermediate values
- We get those values for all positions t in one sweep
- Extend this to message passing (when we have trees)

Minimization

Objective Function:

$$-\log p(\theta|\,\mathbf{X}, Y) = \sum_{i=1}^{m} -\langle\phi(\mathbf{x}_i, y_i), \theta\rangle + g(\theta|\,\mathbf{x}_i) + \frac{1}{2\sigma^2}\|\theta\|^2 +$$

$$\partial_\theta -\log p(\theta|\,\mathbf{X}, Y) = \sum_{i=1}^{m} -\phi(\mathbf{x}_i, y_i) + \mathbb{E}\left[\phi(\mathbf{x}_i, y_i)|\,\mathbf{x}_i\right] + \frac{1}{\sigma^2}\theta$$

We only need $\mathbb{E}\left[\phi_{xy}(\mathbf{x}_{it}, y_{it})|\,\mathbf{x}_i\right]$ and $\mathbb{E}\left[\phi_{yy}(y_{it}, y_{i(t+1)})|\,\mathbf{x}_i\right]$.

Kernel Trick

- Conditional expectations $\Phi(\mathbf{x}_{it}, y_{it})$ pain to compute
- Kernel trick works

$$\langle\phi_{xy}(\mathbf{x}'_t, y'_t), \mathbb{E}\left[\phi_{xy}(\mathbf{x}_t, y_t)|\,\mathbf{x}\right]\rangle = \mathbf{E}\left[k((\mathbf{x}'_t, y'_t), (\mathbf{x}_t, y_t)|\,\mathbf{x}\right]$$

- Get $p(y_t|\,\mathbf{x}, \theta)$, $p(y_t, y_{t+1}|\,\mathbf{x}, \theta)$ via dynamic programming

Subspace Representer Theorem

Representer Theorem:

Solutions of the MAP problem are given by

$$\theta \in \text{span}\{\phi(\mathbf{x}_i, y) \text{ for all } y \in \mathcal{Y} \text{ and } 1 \leq i \leq n\}$$

Big Problem:

$|\mathcal{Y}|$ could be huge, e.g. for sequence annotation 2^n.

Solution:

- Exploit decomposition of $\phi(\mathbf{x}, y)$ into sufficient statistics on cliques.
- Restriction of $\mathcal{Y}$ to cliques is much smaller.

$$\theta_c \in \text{span}\{\phi_c(\mathbf{x}_{ci}, y_c) \text{ for all } y_c \in \mathcal{Y}_c \text{ and } 1 \leq i \leq n\}$$

Rather than 2^n we now get $2^{|c|}$.

Take Home Message

- Exponential families are universal density estimators
- Log partition functions are difficult to compute
- Exponential families + Kernels $\Longrightarrow$ machine learning on steroids
- Gaussian Processes are MAP estimators in feature space
- Novelty detection is density estimation in feature space
- Support Vector Machines are also density estimators
- Margins are the same as odds ratios
- Graphical models and kernels can be married
- Many interesting optimization problems

Questions?

Shameless Plug

- We are hiring at NICTA
- PhD, postdoc and visiting positions are availabe
- Talk to me for more details
  - SVN.Vishwanathan@nicta.com.au