Exponential Families
Classification and Novelty Detection

S.V.N. “Vishy” Vishwanathan

vishy@axiom.anu.edu.au

National ICT of Australia
and
Australian National University

Thanks to Alex Smola, Thomas Hofmann and Stéphane Canu

Overview

- Review of Exponential Family
- Log Partition Function
- Maximum Likelihood Estimation
- MAP Estimation
- Conditional Densities
- Gaussian Processes and the Normal Prior
- Novelty Detection
- Large Margin Classifiers

Exponential Family

Basic Equation:

- We will model densities by

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

Why Exponential Families:

- Dense in space of densities
- The vector $\phi(\mathbf{x})$ closely related to kernels
- We can use $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ where $\mathcal{H}$ is a RKHS
- Conditional models are easy to derive
- Close connections to graphical models

Where is the Catch:

- Statisticians work with explicitly parameterized $\phi(\mathbf{x})$
- The log-partition function is difficult to compute

Log-Partition Function

Basic Equation:

- Some algebra gives us

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

- Computing this integral over $\mathcal{X}$ is painful

Moment Generating Function:

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

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

Other Properties:

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

Universal Density Estimators

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

Conditional Models

Why Condition:

- 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 $\mathbf{x}$

The Answer:

- By Bayes rule we know

$$p(y|\,\mathbf{x};\theta) = \frac{p(\mathbf{x}, y;\theta)}{p(\mathbf{x};\theta)}$$

The Exponential Family:

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

$$\begin{aligned} p(y|\,\mathbf{x};\theta) &= \exp(\langle\phi(\mathbf{x}, y), \theta\rangle - g(\theta|\,\mathbf{x})) \\ g(\theta|\,\mathbf{x}) &= \log \int_{\mathrm{y}} \exp(\langle\phi(\mathbf{x}, \bar{\mathbf{y}}), \theta\rangle \, d\,\bar{\mathbf{y}} \end{aligned}$$

Max-Likelihood Estimation

Basic Idea:

- Given iid data $\mathbf{X} = \{(\mathbf{x}_i, y_i)\}$
- Data drawn from a conditional exponential density
- Find the θ which maximizes $p(y|\,\mathbf{X};\theta)$

The Model:

- By iid assumption

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

The Solution:

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

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

MAP Estimation

Basic Idea:

- We assume that θ is a random variable
- Also assume a prior (belief) over θ
- Now the data updates our belief about the prior

The Normal Prior:

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

$$p(\theta | \mathbf{X}, y) \propto p(y | \mathbf{X}; \theta) p(\theta)$$

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

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 is given by

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

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

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

- By the representer theorem

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

Observations:

- Convex Optimization problem
- If $|\,\mathcal{Y}\,|$ is large we are in trouble
- In case of binary classification we use $\phi(\mathbf{x}, y) = y\phi(\mathbf{x})$

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

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

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

- 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

$$R(\mathbf{x}, y, \theta) = \min_{y \neq y'} \exp(\langle \phi(\mathbf{x}, y) - \phi(\mathbf{x}, y'), \theta \rangle)$$

- 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

Kernel Perceptron

Algorithm:

- Arguably the simplest algorithm in machine learning!
- Maintains a weight vector θ
- Given $(\mathbf{x}_i, y_i)$
  - If $y_i \langle \phi(\mathbf{x}_i), \theta \rangle \geq 0$ do nothing
  - Else $\theta = \theta + y_i \phi(\mathbf{x}_i)$
- Notice how $\theta = \sum_j y_j \phi(\mathbf{x}_j)$

Novikoff's Theorem:

- Given $S = \{(\mathbf{x}_i, y_i\}$ non-trivial
- Let $R = \max_i ||\phi(\mathbf{x}_i)||$ be the radius of the samples
- $\exists\ \theta^*$ such that $||\theta^*|| = 1$ and $y_i \langle \theta^*, \phi(\mathbf{x}_i) \rangle \geq \gamma > 0$
- The kernel perceptron makes at most $(R/\gamma)^2$ mistakes

Odds Ratio As Margin

The Question:

- Consider $\mathcal{Y} = \{1, 2, 3\}$
- Suppose we know $p(1|\,\mathbf{x}_i) = 0.4$
- Can we conclude that $\mathbf{y}_i = 1$?

The Answer:

- No! because we might have $p(2|\,\mathbf{x}_i) = 0.5$
- We want the *true* class probability to be peaked

The Consequences:

- This is equivalent to maximizing the log odds ratio
- Using a normal prior on θ we can solve

$$\min \frac{1}{2}||\theta||^2$$
$$\text{s.t.} \log R(\mathbf{x}_i, y_i, \theta) \geq 1$$

Recovering SVM's

The Problem:

- We solve

$$\min \tfrac{1}{2}||\theta||^2$$
$$\text{s.t.} \log R(\mathbf{x}_i, y_i, \theta) \geq 1$$

- Recall that

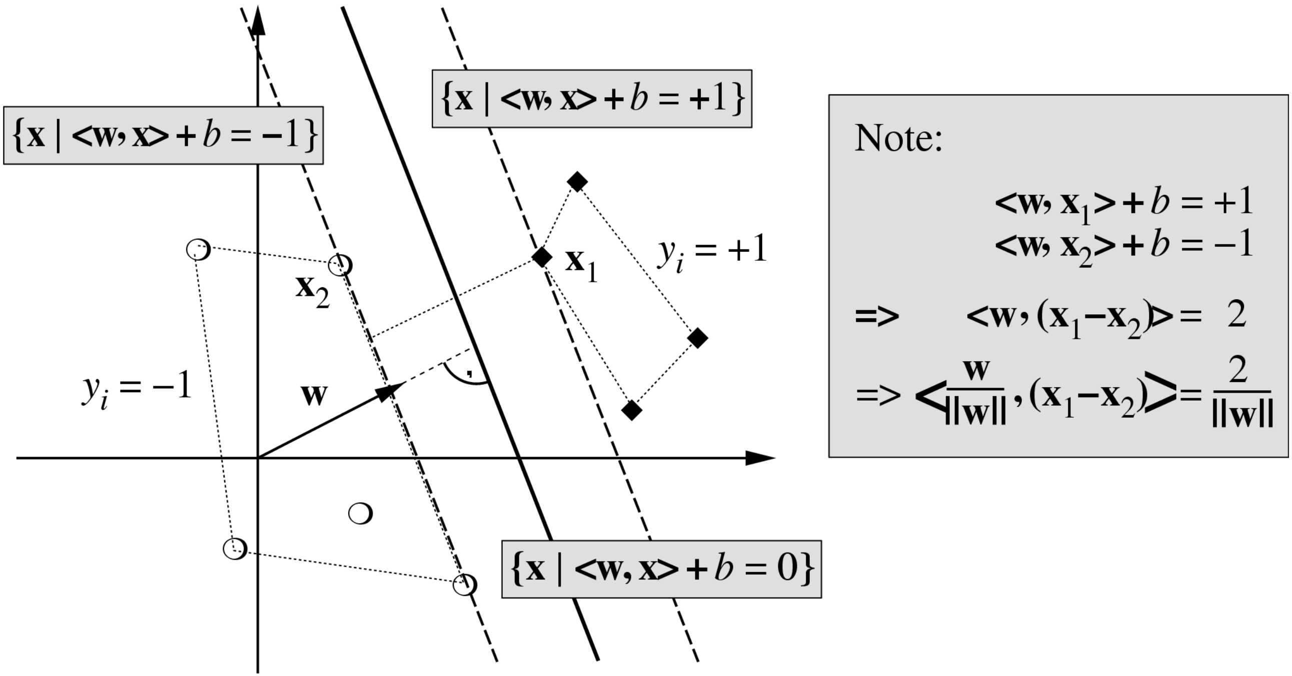
$$\log R(\mathbf{x}, y, \theta) = \max_{y \neq y'} \langle \phi(\mathbf{x}, y) - \phi(\mathbf{x}, y'), \theta \rangle$$

The Binary Case:

- Again recall that $\phi(\mathbf{x}, y) = y\phi(\mathbf{x})$ in the binary case
- Hence $R(\mathbf{x}, y, \theta) = 2\langle \phi(\mathbf{x}), \theta \rangle$
- The equivalent optimization problem is

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

Optimal Separating Hyperplane



Minimize $\frac{1}{2}\|w\|^2$ subject to $y_i(\langle w, x_i\rangle + b) \geq 1$ for all i

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

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

Upper Bound on Error:

- If we define

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

then

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

Another Formulation

Slack Variables:

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

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

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 $\implies$ harder problem to solve
- Solution might not be sparse!

Questions?