Monday, November 3, 2014 3:08 PM

Conjugate Priors
Bayesian Linear Regression
Type II Maximum Likelihood

Admin:
- Project Proposal due
- A6 due friday (in tutorial or mailbox in ICICS 201)
  * don't forget taking an assignment to grade
- Midterm Monday (in class)
  * Study Guide on Piazza
  * you are allowed 2 pages of notes
  * is tutorial this week (Q&A)
  * Auditors: dont come to class! X836 3pm: Dimitri Bertsekas

Learning Principles

	Train	Predict
Maximum Likelihood	$\hat{h} = \underset{h}{\operatorname{argmax}}\ p(D\mid h)$	$p(\hat{D}\mid\hat{h})$
MAP, posterior mode	$\hat{h} = \underset{h}{\operatorname{argmax}}\ p(D\mid\lambda) \propto p(D\mid h)\,p(h\mid\lambda)$ (λ: parameter in prior)	$p(\hat{D}\mid\hat{h})$
Bayesian	$p(h\mid D,\lambda) = \frac{p(D\mid h)\,p(h\mid\lambda)}{p(D\mid\lambda)}$	$p(\hat{D}\mid D,\lambda) = \int p(\hat{D}\mid D,\lambda)\,p(h\mid D,\lambda)\,dh$ or $p(\hat{D}\mid h_{mean})$ → minimizing L_2 risk or $p(\hat{D}\mid h_{median})$ → minimizing absolute risk

Last time: beta ∝ (bernoulli)(beta)
(posterior) ∝ (likelihood)(prior)

Conjugate Priors

Beta prior is conjugate to bernoulli likelihood
- with these choices, prior and posterior are from same family (distribution family)
- makes the integrals "easy"

Likelihood	Conjugate Prior
Bernoulli, Binomial	Beta
Multinoulli, Multinomial	Dirichlet
Gaussian $p(x\mid\mu)$	Gaussian
Gaussian $p(x\mid\sigma^2)$	Inverse Gamma (inverse Chi-squared is a special case)
Gaussian $p(x\mid\mu,\sigma^2)$	Normal Inverse Gamma
Gaussian $p(x\mid\Sigma)$ (covariance matrix)	Inverse Wishart
Gaussian $p(x\mid\mu,\Sigma)$	Normal Inverse Wishart

posterior predictive is student 't' (heavier tailed than Gaussian/Laplace)

$$\iint \underbrace{N(x\mid\mu,\Sigma)}_{\text{normal}}\ \underbrace{NIW(\mu,\Sigma\mid m,k,v,s)}_{\text{normal inverse wishart}}\,d\mu\,d\Sigma$$

* If you don't have conjugate priors:
  - can use mixture of conjugate priors
  - numerical integration (low dimensions)
  - variational methods
    laplace, mean field, variational Bayes, loopy belief propagation, expectation propagation
  - Monte carlo
    gibs sampling, metropolis-hastings, sequential monte carlo

When do we have conjugate priors / they exist?

Exponential Family

θ: parameter
x: data

$$p(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp(\theta^T \phi(x))$$

$h(x)$: scaling constant (usually 1)
$\phi(x)$: "sufficient statistics"
$Z(\theta)$: "partition function" – to sum up to 1

$$\propto \exp(\theta^T \phi(x))[h(x)]$$

$$A(\theta) = \log Z(\theta) \quad \text{"cumulant function"}$$

$$\frac{dA}{d\theta} = E[\phi(x)]$$

$$\frac{d^2A}{d\theta} = \text{var}[\phi(x)]$$

assume $h(x) = 1$

$$NLL(\theta) = -\theta^T \phi(x) + \underbrace{\log Z(\theta)}_{\text{log sum exp}} + \text{const}$$

$\Rightarrow$ convex

$$\nabla NLL(\theta) = -\phi(x) + E[\phi(x)]$$

At MLE: $\phi(x) = E[\phi(x)]$ "moment matching"

fenchel dual: you'll see entropy

Bayesian Linear Regression

$$f(x_i) = w^T x_i$$

$$y = f(x_i) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

$$y_i \sim N(f(x_i), \sigma^2)$$

$$y_i \sim N(w^T x_i, \sigma^2), \quad w_j \sim N(0, \lambda^{-1})$$

← midterm! maybe 1,

$$p(w|X,y) \propto \exp\left(\frac{1}{2\sigma^2}(y - Xw)^T(y - Xw)\right)\exp\left(-\frac{\lambda}{2}\|w\|^2\right)$$

$$\propto \exp\left(-\tfrac{1}{2}(w - \bar{w})^T \underbrace{\left(\tfrac{1}{\sigma^2}X^TX + \lambda I\right)}_{\text{"A"}}{}^{-1}(w - \bar{w})\right)$$

$$\bar{w} = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}X^TX + \lambda I\right)^{-1} X^T y$$

$$w|X,y \sim N\left(\frac{1}{\sigma^2}A^{-1}X^Ty, A^{-1}\right)$$

$\hat{f}|\hat{x}, X, y \Rightarrow \sim N(\cdots)$

Back to Learning Principles:

$p(D|\lambda)$

marginal likelyhood, "evidence"

Type II Maximum Likelyhood
"empirical Bayes"

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\ p(D|\lambda) = \int p(D|h)\,p(h|\lambda)\,dh$$

- Automatic Relevance Determination (ARD)
  "Sparse Bayesian Learning"

$$y_i \sim N(w^T x_i, \theta^2) \qquad w_j \sim N(\emptyset, \lambda_j^{-1})$$

$\lambda_j \to \infty$ for irrelevant features

- Relevance Vector machine
  - RBF basis
  - do ARD

Type II MAP
"hyper-prior"
"hierachical bayesian model"

$$\hat{\lambda} \underset{\lambda}{\operatorname{argmax}}\ p(D|\lambda)\,p(\lambda)$$

"Very Bayesian"

$$p(h, \lambda | D) \propto p(D|h)\,p(h|\lambda)\,p(\lambda)$$

Stochastic Processes

Infinite set of random variables $\{x_i\}$

Gaussian Processes

- Stochastic Process where all finite combinations have a gaussian distribution.
- completely specified by:
  mean function : $m(x) = E[f(x)]$
  covariance function: $k(x,x') = E[(m(x) - f(x))(m(x) - f(x))^T]$

$$f(x) \sim GP(m(x), k(x,x'))$$

Example:

$$f(x) = \phi(x)^T w, \qquad w \sim N(\emptyset, \Sigma)$$

$$E_w[f(x)] = \phi(x)^T \underbrace{E[w]}_{0} = 0$$
$$E_w[f(x) f(x')^T] = \phi(x)^T E[w w^T]\,\phi(x)$$
$$= \phi(x)\,\Sigma\,\phi(x')$$

- Non-parametric Bayes
  - Dirichlet process (stick-breaking, chinese restaurant)
  - Indian Buffet process
  - Polya-urn process