CS340 Machine learning
Final review
Covered in midterm review

- I – Basics:
  - Statistics (MLE, posteriors, Bayes factors, model selection, etc)
  - Info theory
  - Decision theory
• II- Models
  – Generative vs discriminative
  – Naïve Bayes
  – MVN
  – Markov chains
  – DGMs, including expert systems
  – UGMs, including Ising models

• III – Algorithms
  – Gibbs sampling
Unconditional density models

Eg $x \sim \text{bernoulli}, \theta \sim \text{beta}$
$x \sim \text{multinomial}, \theta \sim \text{dir}$
$x \sim \text{multinomial}, \theta \sim \text{mixture of dir}$
$x \sim \text{gaussian}, \theta = (\mu, \lambda) \sim \text{NormalGamma}$
$x \sim \text{MVN}, \theta = (\mu, \Lambda) \sim \text{NormalWishart}$
Generative vs discriminative models

Generative y->x

\[ p(x, y|\pi, \theta) = p(y|\pi)p(x|y, \theta) \]

P(x|y) = class conditional density
Eg  fully factored (naïve Bayes)
    Markov chain
    full covariance Gaussian

Discriminative x->y

\[ p(y|x, w) \]
Naïve Bayes

\[ p(Y = c | x, \theta, \pi) \propto \exp \left[ \log \pi_c + \sum_i I(x_i = 1) \log \theta_{ic} + I(x_i = 0) \log (1 - \theta_{ic}) \right] \]

\[ x' = [1, I(x_1 = 1), I(x_1 = 0), \ldots, I(x_d = 1), I(x_d = 0)] \]

\[ \beta_c = [\log \pi_c, \log \theta_{1c}, \log (1 - \theta_{1c}), \ldots, \log \theta_{dc}, \log (1 - \theta_{dc})] \]

\[ p(Y = c | x, \beta) = \frac{\exp[\beta^T_c x']}{\sum_{c'} \exp[\beta^T_{c'} x']} \]

Becomes sigmoid in 2-class case

\[ P(x_j | Y = c) = \text{bernoulli, gaussian, …} \]

Compute \( p(\theta_j | D) \)

Handle missing data

Log sum exp trick

\[ \beta_c^T x' \]
Multivariate normal

\[ \mathcal{N}(\mathbf{x} | \mu, \Sigma) \overset{\text{def}}{=} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[ -\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right] \]

MLE: \[ \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \Sigma = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) (x_i - \bar{x})^T \]
Gaussian classifiers

Tied Sigma, many classes

\[ p(Y = c | x) = \frac{\pi_c \exp \left[ -\frac{1}{2}(x - \mu_c)^T \Sigma_c^{-1} (x - \mu_c) \right]}{\sum_{c'} \pi_{c'} \exp \left[ -\frac{1}{2}(x - \mu_{c'})^T \Sigma_{c'}^{-1} (x - \mu_{c'}) \right]} \]

\[ = \frac{\exp \left[ \mu_c^T \Sigma_c^{-1} x - \frac{1}{2} \mu_c^T \Sigma_c^{-1} \mu_c + \log \pi_c \right]}{\sum_{c'} \exp \left[ \mu_{c'}^T \Sigma_{c'}^{-1} x - \frac{1}{2} \mu_{c'}^T \Sigma_{c'}^{-1} \mu_{c'} + \log \pi_{c'} \right]} \]

\[ \theta_c \overset{\text{def}}{=} \left( -\frac{\mu_c^T \Sigma_c^{-1} \mu_c + \log \pi_c}{\Sigma_{-1} \mu_c} \right) = \left( \gamma_c \beta_c \right) \]

\[ p(Y = c | x) = \frac{e^{\theta_c^T x}}{\sum_{c'} e^{\theta_{c'}^T x}} = \frac{e^{\beta_c^T x + \gamma_c}}{\sum_{c'} e^{\beta_{c'}^T x + \gamma_{c'}}} \]
Language Models:
Empirical Bayes on rows of $T$
leads to backoff smoothing

$$T_{ij} = \begin{cases} 
\frac{pG_{ij}}{c_j} + \delta & \text{if } c_j \neq 0 \\
\frac{1}{n} & \text{if } c_j = 0 
\end{cases}$$
\[ p(S = 1|W = 1, R = 1) = \frac{p(S = 1, W = 1, R = 1)}{p(W = 1, R = 1)} = 0.19 \]
Expert systems

Pedigree trees

1  2  3  4  5  6

QMR

570 Diseases
flu heart disease botulism

sex=F WBC count abdomen pain

4075 Symptoms

Noisy-or
Undirected graphical models

\[ p(x) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c) \]
Ising models

\[ p(x, z) = p(z)p(x|z) = \left[ \frac{1}{Z} \prod_{<ij>} \psi_{ij}(z_i, z_j) \right] \left[ \prod_i p(x_i|z_i) \right] \]
Gibbs sampling

1. \( x_1^{s+1} \sim p(x_1|x_2^s, \ldots, x_D^s) \)
2. \( x_2^{s+1} \sim p(x_2|x_1^{s+1}, x_3^s, \ldots, x_D^s) \)
3. \( x_i^{s+1} \sim p(x_i|x_{1:i-1}^{s+1}, x_{i+1:D}^s) \)
4. \( x_D^{s+1} \sim p(x_D|x_1^{s+1}, \ldots, x_{D-1}^{s+1}) \)

Monte Carlo integration

\[
p(X_i|X_{-i}) \propto p(X_i|Pa(X_i)) \prod_{Y_j \in \text{ch}(X_i)} p(Y_j|Pa(Y_j))
\]