Lecture 10:

Parameter Learning for Bayes nets

Kevin Murphy

October 20, 2004
Learning graphical models

- Inference means computing $P(X_i | \theta, G)$
- Structure learning/ model selection = inferring $G$ from data.
- Parameter learning/ estimation = inferring $\theta$ from data.
Parameter learning

• Assume $G$ is known and fixed and is a DAG.
• Goal: estimate $\theta$ from a dataset of $M$ independent, identically distributed (iid) training cases $D = (x^1, \ldots, x^M)$.
• In general, each training case $x^m = (x^m_1, \ldots, x^m_N)$ is a vector of values, one per node. (Think of a database with $M$ rows and $N$ columns.)
• We assume complete observability, i.e., every entry in the database is known (no missing values, no hidden variables).
• Initially we consider learning parameters for a single node.
• Then we consider how to learn parameters for a whole network.
Bayesian parameter estimation

- Bayesians treat the unknown parameters $\theta$ as a random variable, which can be estimated using Bayes rule:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- This crucial equation can be written in words:

  posterior = \frac{likelihood \times prior}{marginal likelihood}

- For iid data, the likelihood is

$$p(D|\theta) = \prod_{m} p(x_m|\theta)$$

- The prior $p(\theta)$ encodes our prior knowledge about the domain.
For iid (exchangeable) data, the likelihood is

\[ p(D|\theta) = \prod_{m} p(x_m|\theta) \]

We can represent this as a Bayes net with \( M \) nodes.

“Plates” provide a more compact representation for repetitive structure, and are very common in Bayesian models.
• “Plates” provide a compact representation for repetitive structure.

• The rules of plates are simple: repeat every structure in a box a number of times given by the integer in the corner of the box (e.g. $N$), updating the plate index variable (e.g. $n$) as you go.

• Duplicate every arrow going into the plate and every arrow leaving the plate by connecting the arrows to each copy of the structure.

• Plates are closely related to probabilistic relational models, and object oriented Bayes nets, which are forms of “syntactic sugar” for parameter tying (sharing).
Frequentist parameter estimation

- Two people with different priors $p(\theta)$ will end up with different estimates $p(\theta|D)$.
- Frequentists dislike this “subjectivity”.
- Frequentists think of the parameter as a fixed, unknown constant, not a random variable.
- Hence they have to come up with different estimators (ways of computing $\theta$ from data), instead of using Bayes’ rule.
- These estimators have different properties, such as being “unbiased”, “minimum variance”, etc.
- A very popular estimator is the maximum likelihood estimator, which is simple and has good statistical properties.
Maximum likelihood estimation

- The log-likelihood is monotonically related to the likelihood:
  \[ \ell(\theta; D) = \log p(D|\theta) = \sum_m \log p(x^m|\theta) \]

- Idea of maximum likelihood estimation (MLE): pick the setting of parameters most likely to have generated the data we saw:
  \[ \hat{\theta}_{ML} = \arg\max_\theta \ell(\theta; D) \]

- Often the MLE overfits the training data, so it is common to maximize a penalized log-likelihood instead:
  \[ \hat{\theta}_{MAP} = \arg\max_\theta \ell(\theta; D) - c(\theta) \]

- This is equivalent to picking the mode of \( P(\theta|D) \), where \( c(\theta) = -\log p(\theta) \), since
  \[ \log p(\theta|D) = \log p(D|\theta) + \log p(\theta) + c \]
Integrate out or Optimize?

• \( \hat{\theta}_{MAP} \) is not Bayesian (even though it uses a prior) since it is a point estimate.

• Consider predicting the future. A Bayesian will integrate out all uncertainty:

\[
p(x_{\text{new}}|X) = \int p(x_{\text{new}}, \theta|X) d\theta \\
= \int p(x_{\text{new}}|\theta, X)p(\theta|X) d\theta \\
\propto \int p(x_{\text{new}}|\theta)p(X|\theta)p(\theta) d\theta
\]

• A frequentist will typically use a “plug-in” estimator such as ML/MAP:

\[
p(x_{\text{new}}|X) = p(x_{\text{new}}|\hat{\theta}), \quad \hat{\theta} = \arg \max_\theta p(X|\theta)
\]
**Frequentist vs Bayesian**

- This is a “theological” war.

- Advantages of Bayesian approach:
  - Mathematically elegant.
  - Works well when amount of data is much less than number of parameters (e.g., one-shot learning).
  - Easy to do incremental (sequential) learning.
  - Can be used for model selection (max likelihood will always pick the most complex model).

- Advantages of frequentist approach:
  - Mathematically/computationally simpler.

- As $|D| \to \infty$, the two approaches become the same:

$$p(\theta|D) \to \delta(\theta, \hat{\theta}_{ML})$$
**Example MLE: Bernoulli Trials**

- We observe $M$ iid coin flips: $\mathcal{D} = H, H, T, H, \ldots$
- Model: $p(H) = \theta$  $p(T) = (1 - \theta)$
- Likelihood:

  $$\ell(\theta; \mathcal{D}) = \log p(\mathcal{D}|\theta) = \log \prod_{m} \theta^{x_m} (1 - \theta)^{1-x_m}$$

  $$= \log \theta \sum_{m} x_m + \log(1 - \theta) \sum_{m} (1 - x_m)$$

  $$\Rightarrow \theta^*_{\text{ML}} = \frac{N_H}{N_H + N_T}$$
Sufficient statistics

• The counts $N_H = \sum_m x^m$ and $N_T = \sum_m (1 - x^m)$ are sufficient statistics of the data $D$.

• In general, $T(X)$ is a sufficient statistic for $X$ if

$$T(x^1) = T(x^2) \Rightarrow L(\theta; x^1) = L(\theta; x^2)$$
**Example: Multinomial**

- We observe $M$ iid die rolls (K-sided): $D=3,1,K,2,\ldots$
- Model: $p(k) = \theta_k$, $\sum_k \theta_k = 1$
- Likelihood (for binary indicators $[x^m = k]$):
  
  $$\ell(\theta; D) = \log p(D|\theta) = \sum_m \log \prod_k \theta^{[x^m = k]}$$
  
  $$= \sum_m \sum_k [x^m = k] \log \theta_k = \sum_k N_k \log \theta_k$$

- We need to maximize this subject to the constraint $\sum_k \theta_k = 1$, so we use a Lagrange multiplier.
Lagrange multipliers

- Constrained cost function:

\[
\tilde{l} = \sum_k N_k \log \theta_k + \lambda \left( 1 - \sum_k \theta_k \right)
\]

- Take derivatives wrt $\theta_k$:

\[
\frac{\partial \tilde{l}}{\partial \theta_k} = \frac{N_k}{\theta_k} - \lambda = 0
\]

\[
N_k = \lambda \theta_k
\]

\[
\sum_k N_k = M = \lambda \sum_k \theta_k = \lambda
\]

\[
\hat{\theta}_{k,ML} = \frac{N_k}{M}
\]

- $\hat{\theta}_{k,ML}$ if the fraction of times $k$ occurs.
Example: Univariate Normal
Example: Univariate Normal

• We observe \( M \) iid real samples: \( D=1.18, -.25, .78, \ldots \)
• Model: \( p(x) = (2\pi\sigma^2)^{-1/2} \exp\{-(x-\mu)^2/2\sigma^2\} \)

Log likelihood:

\[
\ell(\theta; D) = \log p(D|\theta)
= -\frac{M}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum_m \frac{(x^m - \mu)^2}{\sigma^2}
\]

• Take derivatives and set to zero:

\[
\frac{\partial \ell}{\partial \mu} = (1/\sigma^2) \sum_m (x_m - \mu)
\]

\[
\frac{\partial \ell}{\partial \sigma^2} = -\frac{M}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_m (x_m - \mu)^2
\]

\[
\Rightarrow \mu_{ML} = \frac{1}{M} \sum_m x_m
\]

\[
\sigma^2_{ML} = \frac{1}{M} \sum_m (x_m - \mu_{ML})^2
\]
For a numeric random variable $x$

$$p(x|\eta) = h(x) \exp\{\eta^\top T(x) - A(\eta)\}$$

$$= \frac{1}{Z(\eta)} h(x) \exp\{\eta^\top T(x)\}$$

is an exponential family distribution with natural (canonical) parameter $\eta$.

- Function $T(x)$ is a sufficient statistic.
- Function $A(\eta) = \log Z(\eta)$ is the log normalizer.
- Examples: Bernoulli, multinomial, Gaussian, Poisson, gamma, ...
- A distribution $p(x)$ has finite sufficient statistics (independent of number of data cases) iff it is in the exponential family.
**Multivariate Gaussian Distribution**

- For a continuous vector random variable:
  \[
p(x|\mu, \Sigma) = |2\pi\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2}(x - \mu)^{\top}\Sigma^{-1}(x - \mu) \right\}
  \]

- Exponential family with:
  \[
  \eta = [\Sigma^{-1}\mu; -1/2\Sigma^{-1}]
  \]
  \[
  T(x) = [x; xx^{\top}]
  \]
  \[
  A(\eta) = \log |\Sigma|/2 + \mu^{\top}\Sigma^{-1}\mu/2
  \]
  \[
  h(x) = (2\pi)^{-d/2}
  \]

- Note: a d-dimensional Gaussian is a d+d^2-parameter distribution with a d+d^2-component vector of sufficient statistics (but because of symmetry and positivity, parameters are constrained)
Moments

- We can easily compute moments of any exponential family distribution by taking the derivatives of the log normalizer $A(\eta)$.
- The $q^{th}$ derivative gives the $q^{th}$ centred moment.

\[
\frac{dA(\eta)}{d\eta} = \text{mean} \\
\frac{d^2A(\eta)}{d\eta^2} = \text{variance} \\
\ldots
\]

- When the sufficient statistic is a vector, partial derivatives need to be considered.
Moments

\[
\frac{dA}{d\eta} = \frac{d}{d\eta} \log Z(\eta) = \frac{1}{Z(\eta)} \frac{d}{d\eta} Z(\eta)
\]
\[
= \frac{1}{Z(\eta)} \frac{d}{d\eta} \int h(x) \exp \{\eta T(x)\} dx
\]
\[
= \frac{\int T(x) h(x) \exp \{\eta T(x)\}}{Z(\eta)}
\]
\[
= ET(X)
\]

\[
\frac{d^2 A}{d\eta^2} = Var T(X)
\]
Moment vs canonical parameters

- The moment parameter $\mu$ can be derived from the natural (canonical) parameter

$$\frac{dA}{d\eta} = ET(X) \overset{\text{def}}{=} \mu$$

- Now $A(\eta)$ is convex since

$$\frac{d^2 A}{d\eta^2} = VarT(X) > 0$$

- Hence we can invert the relationship and infer the canonical parameter from the moment parameter:

$$\eta \overset{\text{def}}{=} \psi(\mu)$$
For iid data, the log-likelihood is

\[
\ell(\eta; D) = \log \prod_m h(x^m) \exp \left( \eta^T T(x^m) - A(\eta) \right)
\]

\[
= \left( \sum_m \log h(x^m) \right) - MA(\eta) + \left( \eta^\top \sum_m T(x^m) \right)
\]

Take derivatives and set to zero:

\[
\frac{\partial \ell}{\partial \eta} = \sum_m T(x^m) - M \frac{\partial A(\eta)}{\partial \eta} = 0
\]

\[
\Rightarrow \frac{\partial A(\eta)}{\partial \eta} = \frac{1}{M} \sum_m T(x^m)
\]

\[
\hat{\mu}_{ML} = \frac{1}{M} \sum_m T(x^m)
\]

This amounts to moment matching.

We can infer the canonical parameters using \( \hat{\eta}_{ML} = \psi(\hat{\mu}_{ML}) \)
• If we assume the parameters for each CPD are globally independent, then the log-likelihood function decomposes into a sum of local terms, one per node:

$$\log p(D|\theta) = \log \prod_m \prod_i p(x_i^m|x_{\pi_i}, \theta_i) = \sum_i \sum_m \log p(x_i^m|x_{\pi_i}, \theta_i)$$
**Example: A Directed Model**

- Consider the distribution defined by the DAGM:
  \[
p(x|\theta) = p(x_1|\theta_1)p(x_2|x_1, \theta_2)p(x_3|x_1, \theta_3)p(x_4|x_2, x_3, \theta_4)
\]
- This is exactly like learning four separate small DAGMs, each of which consists of a node and its parents.
MLE for Bayes nets with tabular CPDs

- Assume each CPD is represented as a table (multinomial) where
  \[ \theta_{ijk} \overset{\text{def}}{=} P(X_i = j | X_{\pi_i} = k) \]
- The sufficient statistics are just counts of family configurations
  \[ N_{ijk} \overset{\text{def}}{=} \sum_m I(X^m_i = j, X^m_{\pi_i} = k) \]
- The log-likelihood is
  \[
  \ell = \log \prod_m \prod_{ijk} \theta_{ijk}^{N_{ijk}} \\
  = \sum_m \sum_{ijk} N_{ijk} \log \theta_{ijk}
  \]
- Using a Lagrange multiplier to enforce so \( \sum_j \theta_{ijk} = 1 \) we get
  \[
  \hat{\theta}_{ijk}^{ML} = \frac{N_{ijk}}{\sum_{j'} N_{ij'k}}
  \]
• Consider a time-invariant hidden Markov model (HMM)
  
  – State transition matrix $A(i, j) \overset{\text{def}}{=} P(X_t = j | X_{t-1} = i)$,
  – Discrete observation matrix $B(i, j) \overset{\text{def}}{=} P(Y_t = j | X_t = i)$
  – State prior $\pi(i) \overset{\text{def}}{=} P(X_1 = i)$.

The joint is

$$P(X_{1:T}, Y_{1:T} | \theta) = P(X_1 | \pi) \prod_{t=2}^{T} P(X_t | X_{t-1}, A) \prod_{t=1}^{T} P(Y_t | X_t; B)$$
Learning a fully observed HMM

- The log-likelihood is

\[
\ell(\theta; D) = \sum_m \log P(X_1 = x_1^m | \pi) + \sum_{t=2}^T P(X_t = x_t^m | X_{t-1} = x_{t-1}^m, A) + \sum_{t=1}^T P(Y_t = y_t^m | X_t = x_t^m, B)
\]

- We can optimize each parameter \((A, B, \pi)\) separately.
Learning a Markov chain transition matrix

• Define $A(i, j) = P(X_t = j | X_{t-1} = i)$.
• $A$ is a stochastic matrix: $\sum_j A(i, j) = 1$
• Each row of $A$ is multinomial distribution.
• So MLE is the fraction of transitions from $i$ to $j$

$$\hat{A}_{ML}(i, j) = \frac{\#i \rightarrow j}{\sum_k \#i \rightarrow k} = \frac{\sum_m \sum_{t=2}^{T} I(X_{t-1}^m = i, X_t^m = j)}{\sum_m \sum_{t=2}^{T} I(X_{t-1}^m = i)}$$

• If the states $X_t$ represent words, this is called a bigram language model.
• Note that $\hat{A}_{ML}(i, j) = 0$ if the particular $i, j$ pair did not occur in the training data; this is called the sparse data problem.
• We will solve this later using a prior.
CPDs for continuous nodes

- So far we have considered the case where \( p(y|x, \theta) \) can be represented as a multinomial (table).
- Now we consider the case where some nodes may be continuous.

| X  | Y         | \( p(Y|X) \)                      |
|----|-----------|----------------------------------|
| \( \mathbb{R}^n \) | \( \mathbb{R}^m \) | regression                       |
| \( \mathbb{R}^n \) | \( \{0, 1\} \)   | binary classification             |
| \( \{0, 1\}^n \)   | \( \{0, 1\} \)   | binary classification             |
| \( \mathbb{R}^n \) | \( \{1, \ldots, K\} \) | multiclass classification         |
| \( \{1, \ldots, K\} \) | \( \mathbb{R}^n \) | conditional density modeling      |
Learning a conditional Gaussian

- Consider an HMM with discrete states $X_t$ but continuous observations $y_t \in \mathbb{R}^n$:
  \[ p(y_t|X_t = i) = \mathcal{N}(y_t; \mu_i, \Sigma_i) \]

- The MLE is the sample mean and sample variance of observations associated with each state (use $X_t$ labels to partition the data):
  \[
  \hat{\mu}_{ML}(i) = \frac{\sum_{m,t: X_t^m = i} y_t^m}{\sum_{m,t} y_t^m} = \frac{\sum_m \sum_{t=1}^T I(X_t^m = i) y_t^m}{\sum_m \sum_{t=1}^T y_t^m}
  \]

- Note that the MLE for $\Sigma_i$ for states $i$ with small numbers of observations is $\Sigma_i \to \infty I$.

- We will solve this later using a prior.