Probabilistic Machine Learning
Bayesian Nets, MCMC, and more

Marek Petrik

4/18/2017

Conditional Independence

- Independent random variables

\[ \mathbb{P}[X, Y] = \mathbb{P}[X] \mathbb{P}[Y] \]

- Convenient, but not true often enough
Conditional Independence

- Independent random variables
  \[ \mathbb{P}[X, Y] = \mathbb{P}[X] \mathbb{P}[Y] \]

- Convenient, but not true often enough

- **Conditional** independence
  \[ X \perp Y \mid Z \iff \mathbb{P}[X, Y \mid Z] = \mathbb{P}[X \mid Z] \mathbb{P}[Y \mid Z] \]

- Use conditional independence in machine learning
Dependent but Conditionally Independent

Events with a possibly biased coin:

1. $X$: Your first coin flip is heads
2. $Y$: Your second flip is heads
3. $Z$: Coin is biased
Dependent but Conditionally Independent

Events with a possibly biased coin:
1. $X$: Your first coin flip is heads
2. $Y$: Your second flip is heads
3. $Z$: Coin is biased

- $X$ and $Y$ are not independent
- $X$ and $Y$ are independent given $Z$
Independent but Conditionally Dependent

Is this possible?
Independent but Conditionally Dependent

Is this possible? **Yes!** Events with an unbiased coin:

1. $X$: Your first coin flip is heads
2. $Y$: Your second flip is heads
3. $Z$: The coin flips are the same
Independent but Conditionally Dependent

Is this possible? Yes! Events with an unbiased coin:

1. $X$: Your first coin flip is heads
2. $Y$: Your second flip is heads
3. $Z$: The coin flips are the same

- $X$ and $Y$ are independent
- $X$ and $Y$ are not independent given $Z$
Conditional Independence in Machine Learning

- Linear regression
Conditional Independence in Machine Learning

- Linear regression
- LDA
Conditional Independence in Machine Learning

- Linear regression
- LDA
- Naive Bayes
Directed Graphical Models

- Represent complex structure of conditional independence
Directed Graphical Models

- Represent complex structure of conditional independence
- Node is independent of all predecessors **conditional** on parent value

\[ x_s \perp x_{\text{pred}(s) \setminus \text{pa}(s)} \mid x_{\text{pa}(s)} \]
Undirected Graphical Models

- Another (different) representation of conditional independence

- Markov Random Fields
Naive Bayes Model

Closely related to QDA and LDA
Naive Bayes Model

- Chain rule

\[
P[x_1, x_2, x_3] = P[x_1]P[x_2|x_1]P[x_3|x_1, x_2]
\]

- Probability

\[
P[x, y] = P[y] \prod_{j=1}^{D} P[x_j|y]
\]
Why Bother with Conditional Independence?
Why Bother with Conditional Independence?

- Reduces number of parameters
Why Bother with Conditional Independence?

- Reduces number of parameters
- Reduces bias or variance?
Markov Chain

- 1st order Markov chain:

- 2nd order Markov chain:
Uses of Markov Chains

- Time series prediction
- Simulation of stochastic systems
- Inference in Bayesian nets and models
- Many others …
Hidden Markov Models

Used for:

- Speech and language recognition
- Time series prediction
- **Kalman filter**: version with normal distributions used in GPS’s
Inference

- Inference of hidden variables ($y$)

$$P[y|x, \theta] = \frac{P[y, x|\theta]}{P[x|\theta]}$$

- Eliminating nuisance variables (e.g. $x_1$ is not observed)

$$P[y|x_2, \theta] = \sum_{x_1} P[y, x_1|x_2, \theta]$$

- What is inference in linear regression?
Learning

- Computing conditional probabilities $\theta$
- Approaches:
  1. Maximum A Posteriori (MAP)

$$\arg\max_{\theta} \log P[\theta|x] = \arg\max_{\theta} \left( \log P[x|\theta] + \log P[\theta] \right)$$
Learning

- Computing conditional probabilities $\theta$
- Approaches:
  1. Maximum A Posteriori (MAP)

\[
\arg \max_{\theta} \log P[\theta|x] = \arg \max_{\theta} \left( \log P[x|\theta] + \log P[\theta] \right)
\]

2. Inference!
   - Infer distribution of $\theta$ given $x$
   - Return mode, median, mean, or anything appropriate
Learning

- Computing conditional probabilities $\theta$
  - Approaches:
    1. Maximum A Posteriori (MAP)
      \[
      \arg\max_{\theta} \log P[\theta|x] = \arg\max_{\theta} \left( \log P[x|\theta] + \log P[\theta] \right)
      \]
    2. Inference!
      - Infer distribution of $\theta$ given $x$
      - Return mode, median, mean, or anything appropriate
  - Fixed effects vs random effects (mixed effects models)
Inference in Practice

- Precise inference is often impossible
- Variational inference: approximate models
- Markov Chain Monte Carlo (MCMC):
  - Gibbs samples
  - Metropolis Hastings
  - Others
Probabilistic Modeling Languages

- Simple framework to describe a Bayesian model
- Inference with MCMC and parameter search
- Popular frameworks:
  - JAGS
  - BUGS, WinBUGS, OpenBUGS
  - Stan
- Examples:
  - Linear regression
  - Ridge regression
  - Lasso