Introduction to Machine Learning – Generative Approach

The 7th KIAS CAC Summer School
2016. 6. 27 (Mon.)

Yung-Kyun Noh

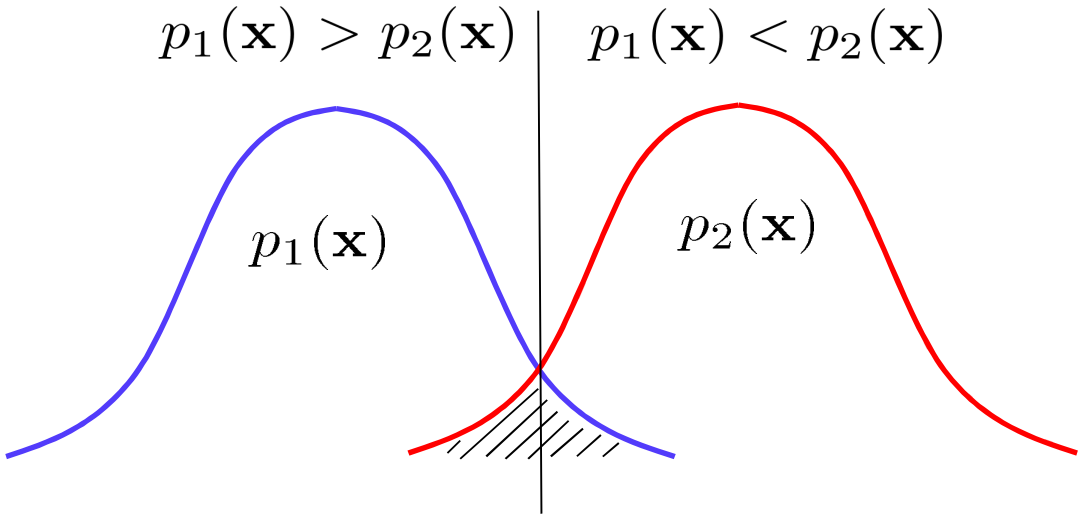
Seoul National University

Overview

- Motivation from the classification perspective
- Independence and model complexity
- How to incorporate independency while the model structure is kept intact as much as possible
- Directed graphical model and undirected graphical model

Probabilistic Assumption and Bayes Classification

- Bayes classification produces theoretical minimum error

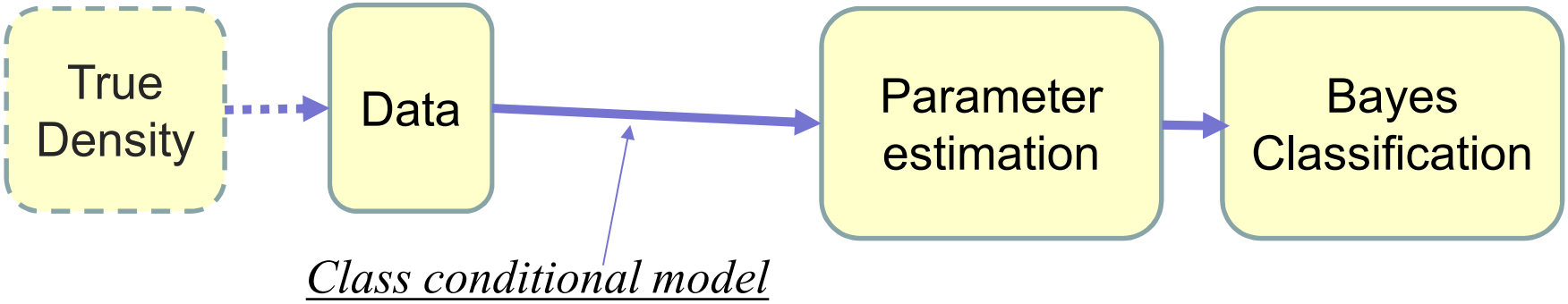


➔ Error:

$$\frac{1}{2}\int \min[p_1, p_2]\, d\mathbf{x}$$

Generative Model

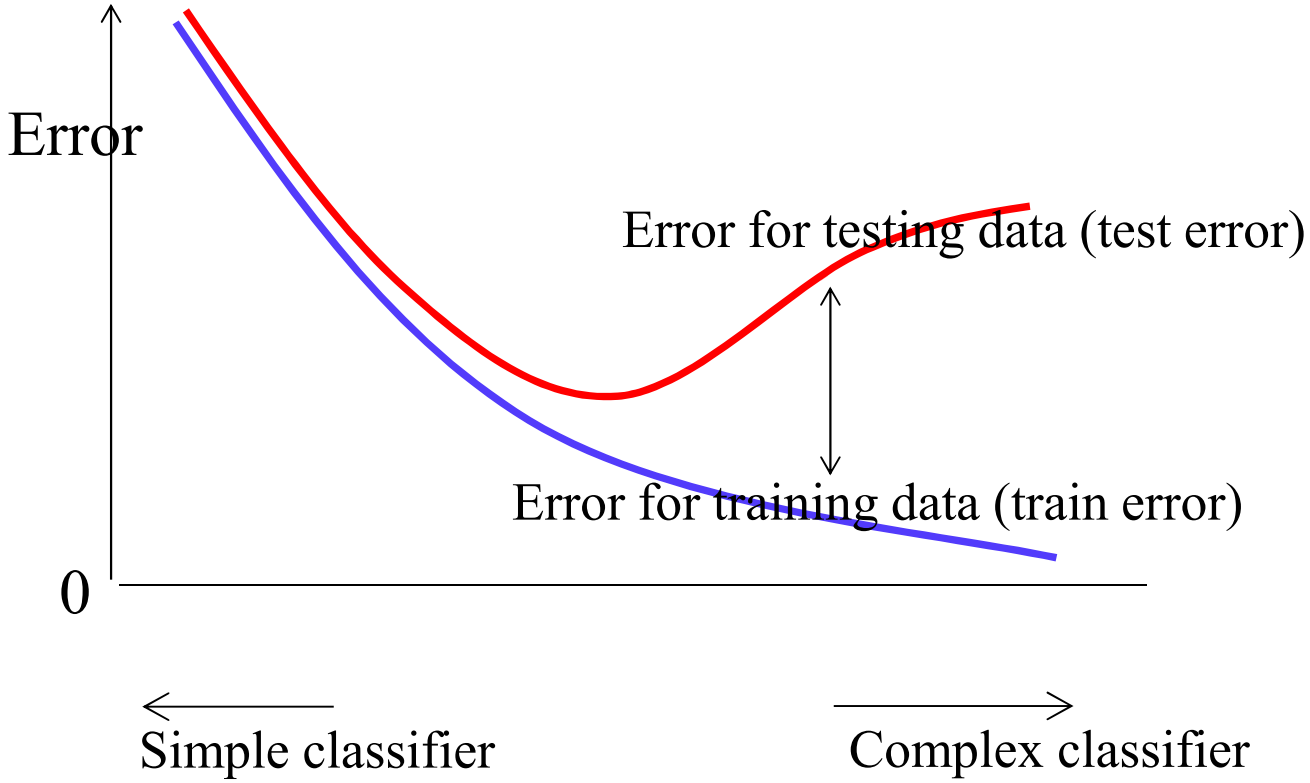
- Generative method for classification
  - Perform the Bayes classification
  - But we now use model instead of true underlying distribution



  - Bayes classification now uses model with estimated parameters, which is not true density, but is now considered as true density.

When do Algorithms Malfunction?

- Generalization and Overfitting



Complexity of Algorithms

- Statistical learning theory for classification
  - with probability at least 1 – δ
  - h: VC-dimension, n: number of data
  - *R(g)*: true risk of function *g*, $R_n(g)$: empirical risk of function *g*

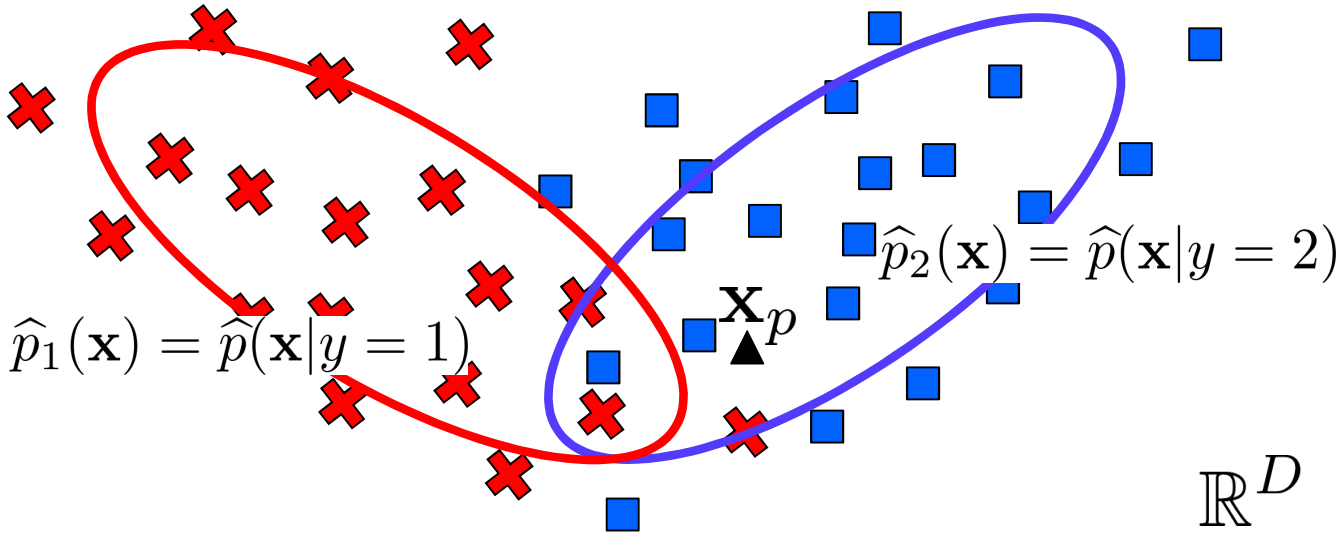
$$R(g) \leq R_n(g) + 2\sqrt{2\frac{h\log(2en/h) + \log(2/\delta)}{n}}$$

- Linear classifier
  - VC-dim = dimensionality + 1

- Generative model
  - Number of parameters

Model + Estimated Parameters

- Ex. Gaussian model



$$\widehat{p}_1(\mathbf{x}_p) \geq \widehat{p}_2(\mathbf{x}_p) \rightarrow y_p = 1$$
$$\widehat{p}_1(\mathbf{x}_p) < \widehat{p}_2(\mathbf{x}_p) \rightarrow y_p = 2$$

Model + Estimated Parameters

- Ex. Gaussian model

$$\widehat{p}(\mathbf{x}) = \mathcal{N}(\widehat{\mu}, \widehat{\Sigma}) \qquad \mathbf{x}, \widehat{\mu} \in \Re^D, \ \widehat{\Sigma} \in \Re^{D \times D}$$

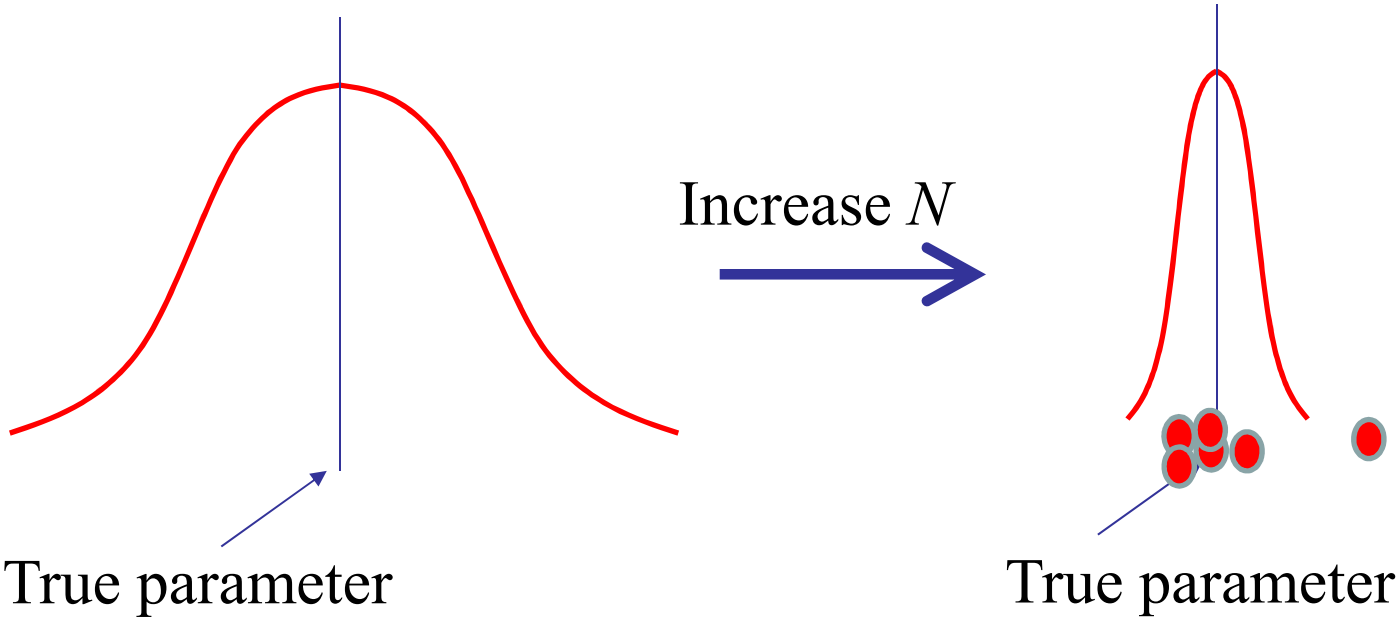
$$= \frac{1}{\sqrt{2\pi}^D |\widehat{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mathbf{x} - \widehat{\mu})^\top\right)$$

- Unbiased estimators

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \qquad \widehat{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N} (\mathbf{x}_i - \widehat{\mu})(\mathbf{x}_i - \widehat{\mu})^\top$$

Unbiased Estimation

- Consistency



Theory: Cramer-Rao bound
Minimum variance of covariance estimator: σ^2/n

Covariance Estimation

- In high-dimensional space

$$\Sigma_{D\times D} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1D} \\ \sigma_{21} & \sigma_2^2 & & & \vdots \\ \vdots & & \ddots & \sigma_{ij} & \\ \vdots & & \sigma_{ij} & & \\ \sigma_{D1} & \cdots & & & \sigma_D^2 \end{pmatrix}$$

(D + 1)D/2 number of parameters for covariances

Number of Parameters

- D = 1000

  – Number of parameters of a Gaussian:

  1000 + 1001*(1000)/2 = 501,500

Independence

- $p(\mathbf{x}) = p_1(\mathbf{x}_1)p_2(\mathbf{x}_2)$

  $\mathbf{x} \in \mathbb{R}^D, \mathbf{x}_1 \in \mathbb{R}^{D_1}, \mathbf{x}_2 \in \mathbb{R}^{D_2} \qquad D = D_1 + D_2$

- D_1 = 500, D_2 = 500
  - Number of parameters

    500 + 501*(500)/2 + 500 + 501*(500)/2

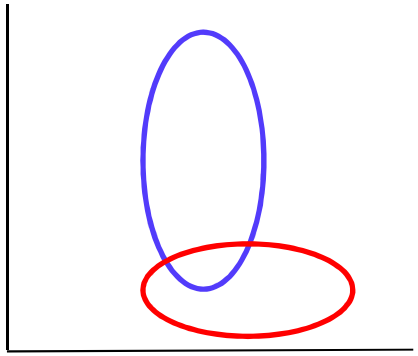
    = 251,500

- Incorporating one independence can reduce the number of parameters into half.

Naïve Bayes As An Extreme Case

- Naïve Bayes

$$p(\mathbf{x}) = \prod_{d=1}^{D} p_d(x_d)$$
$$= p_1(x_1)p_2(x_2)\ldots p_D(x_D)$$



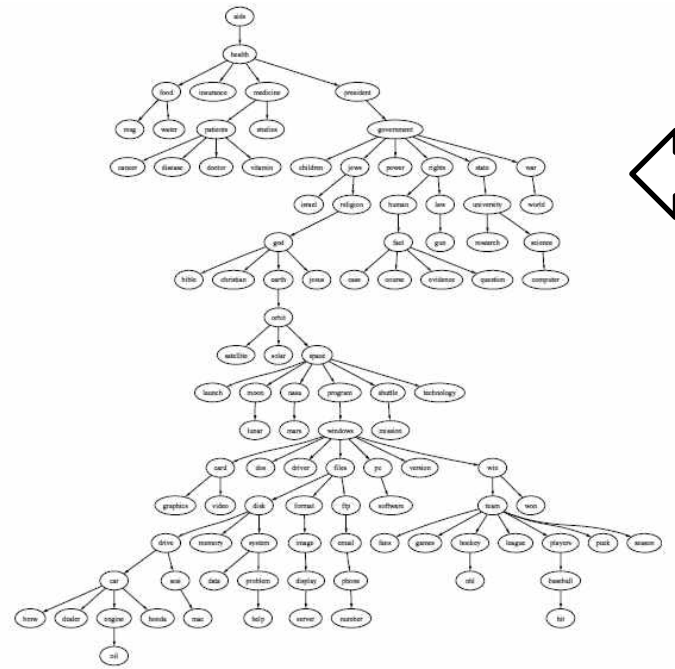
  - Simply ignore every correlation and dependencies between variables

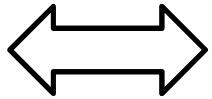
- True decomposition:

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)\ldots p(x_D|x_1,\ldots,x_{D-1})$$

Graphical Models

- We utilize probabilities that are represented by the graph structure. (directed & undirected)

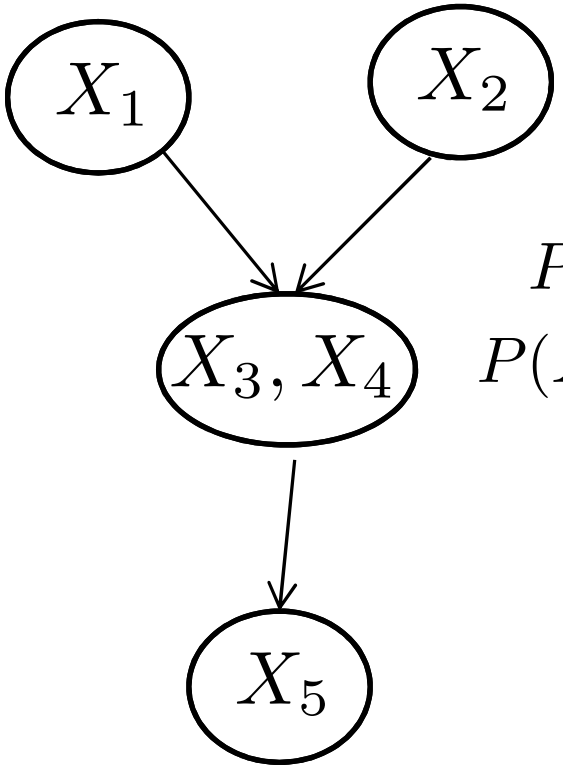




Use probabilisic ***independencies*** and ***conditional independencies*** that can be captured by graph structure

Causality Graph

- Directed Acyclic Graph (DAG)

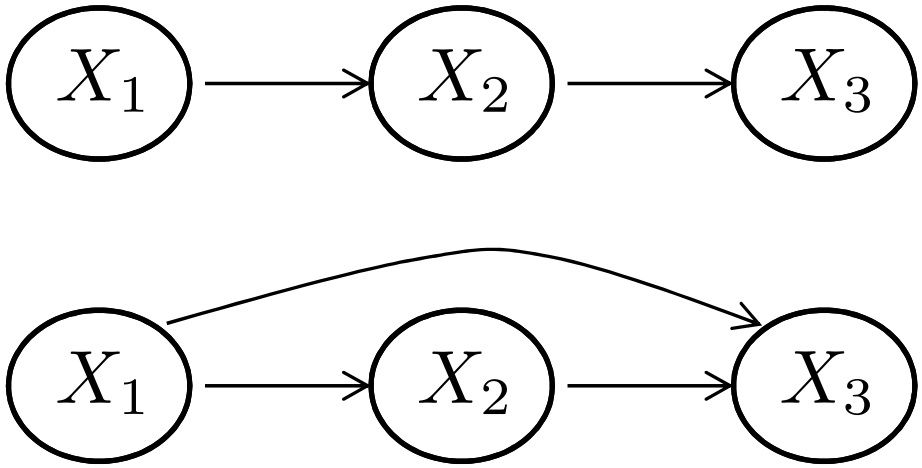


$$P(X_1, X_2, X_3, X_4, X_5) =$$
$$P(X_1)P(X_2)P(X_3, X_4|X_1, X_2)P(X_5|X_3, X_4)$$

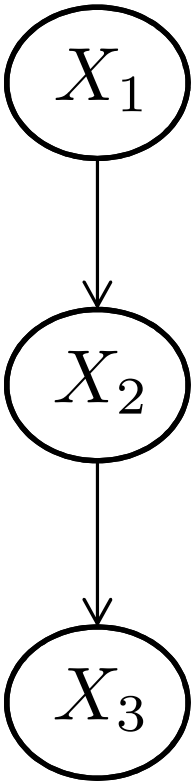
$$P(X_1, \ldots, X_D) = \prod_{i=1}^{D} P(X_i|\mathbf{Pa}_{X_i})$$

Question?

- What is the difference between two graphs?

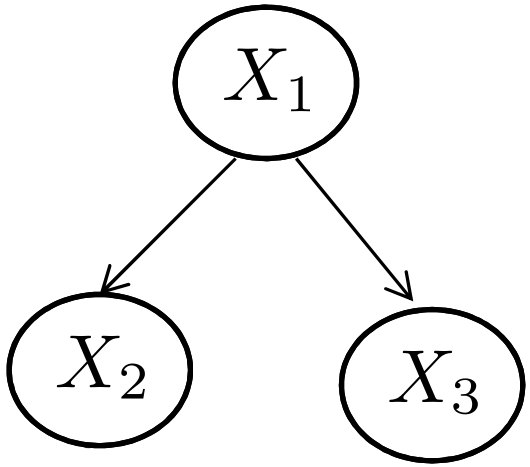


D-Separations



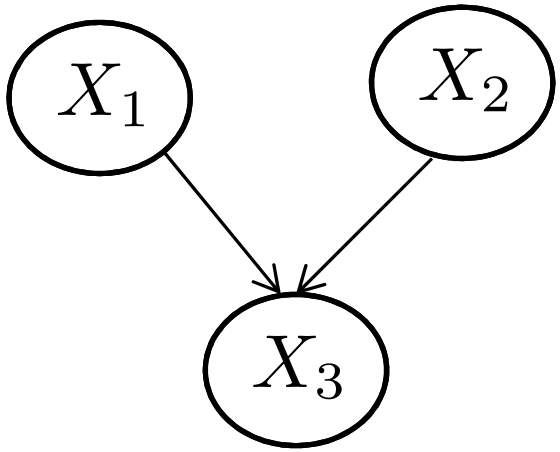
$X_1 \perp\!\!\!\perp X_3 | X_2$

Causal path



$X_2 \perp\!\!\!\perp X_3 | X_1$

Common cause

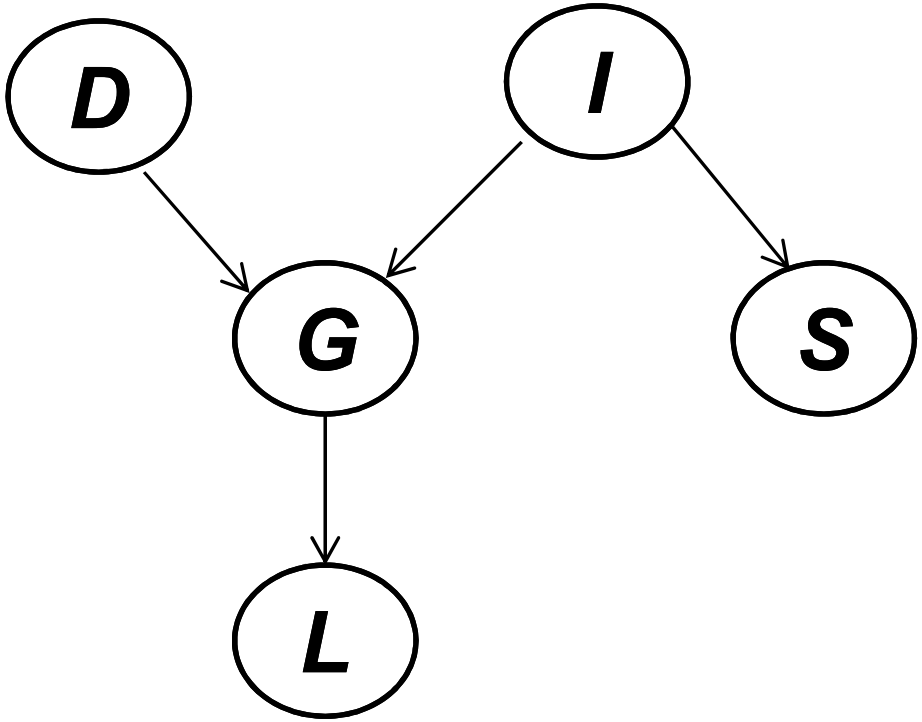


$X_1 \perp\!\!\!\perp X_2$

$X_1 \not\perp\!\!\!\perp X_2 | X_3$

Common effect

Want to get a good reference letter?



D: *Difficulty*
I: *Intelligence*
G: *Grade*
S: *SAT*
L: *Reference Letter*

$$I \perp\!\!\!\perp L|G$$

$$G \perp\!\!\!\perp S|I$$

$$D \perp\!\!\!\perp I$$

$$D \not\perp\!\!\!\perp I|G$$

Infant Rearing

Sleepy

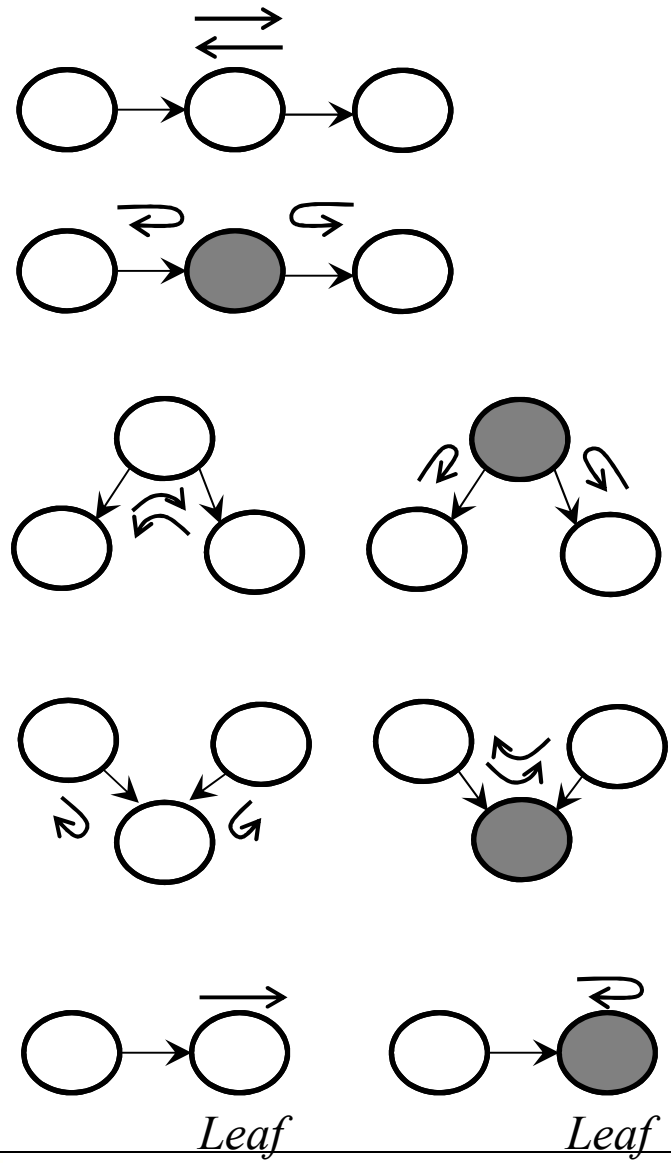
Hungry

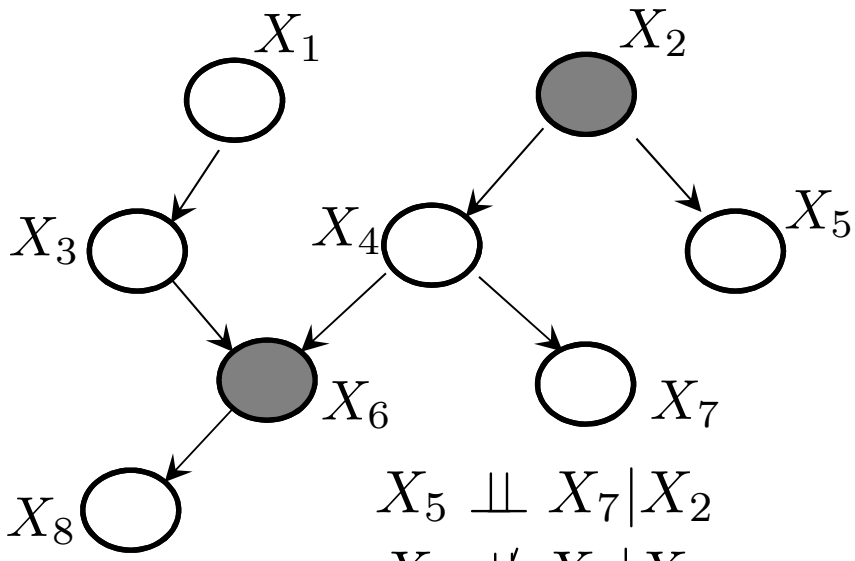
diaper

Fever

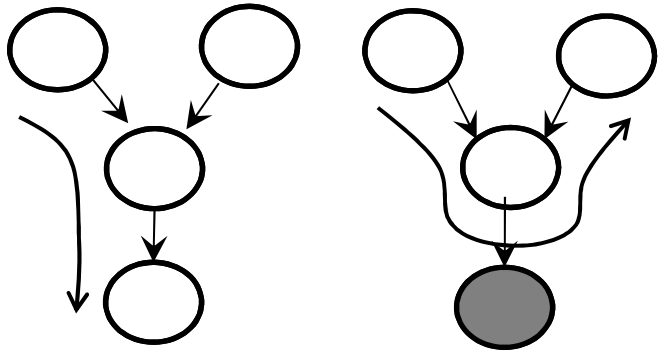
Baby Cry

Bayes Ball Theorem



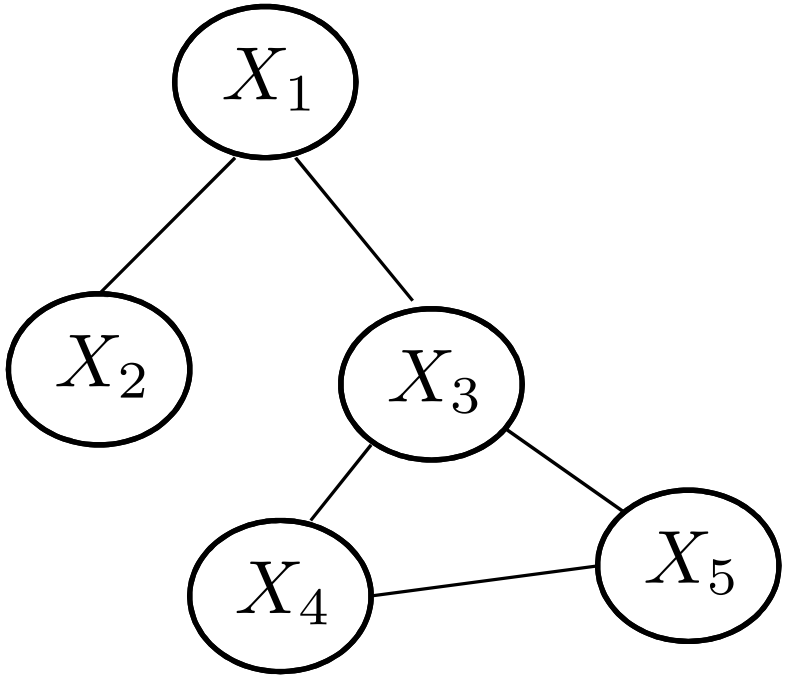


$$X_5 \perp\!\!\!\perp X_7 | X_2$$
$$X_1 \not\perp\!\!\!\perp X_7 | X_6$$
$$X_1 \perp\!\!\!\perp X_8 | X_6$$



Markov Random Field

- Undirected Graph



If there is a direct edge between X_i and X_j:

$$X_i \not\perp\!\!\!\perp X_j | X_{\setminus i,j}$$

If there is no direct edge between X_i and X_j:

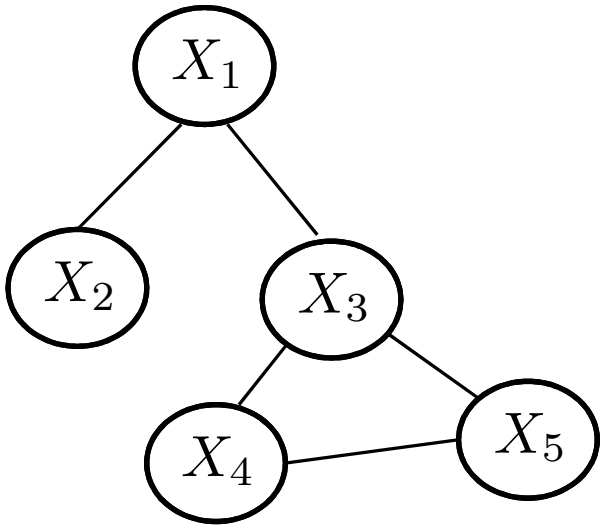
$$X_i \perp\!\!\!\perp X_j | X_{\setminus i,j}$$

$$X_1 \not\perp\!\!\!\perp X_3 | X_2, X_4, X_5$$

$$X_1 \perp\!\!\!\perp X_5 | X_3$$

Joint Distribution

- Product of functions on cliques

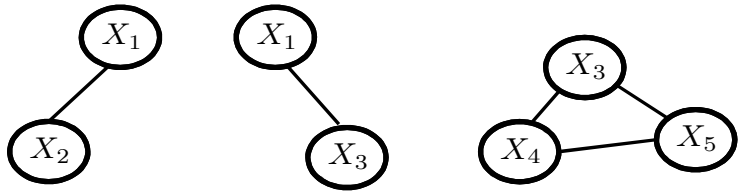


$$P(X_1, X_2, X_3, X_4, X_5) =$$

$$\frac{1}{Z}\ \psi_{1,2}(X_1, X_2)\ \psi_{1,3}(X_1, X_3)\ \psi_{3,4,5}(X_3, X_4, X_5)$$

$$\left(Z = \sum_{X_1, X_2, X_3, X_4, X_5} \psi_{1,2}(X_1, X_2)\psi_{1,3}(X_1, X_3)\psi_{3,4,5}(X_3, X_4, X_5) \right)$$

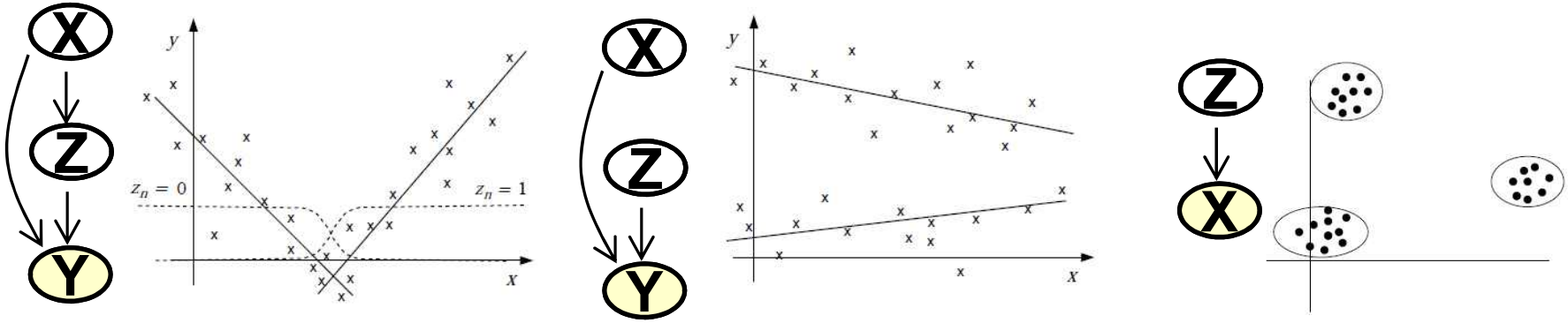
Cliques:



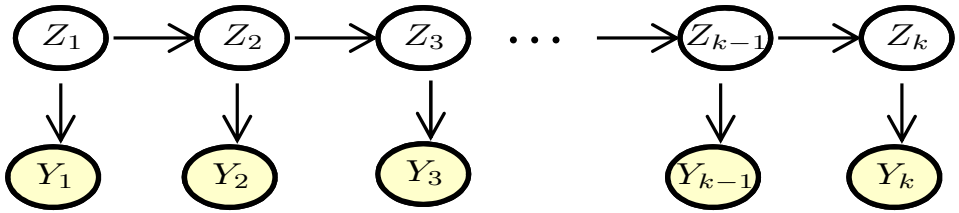
The set of distributions satisfying MRF conditions (Markov random field) = The set of distributions decomposed by cliques (Gibbs random field) (Hammersley-Clifford Theorem)

More Fancy Models

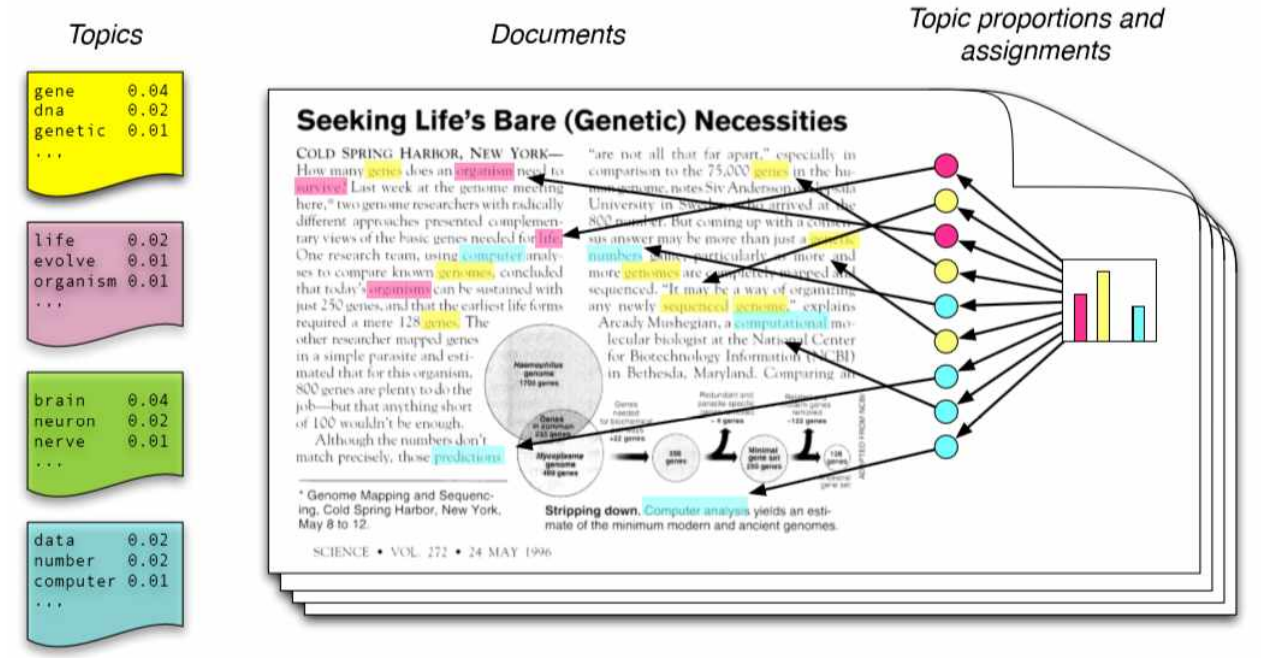
- Latent Variable Model

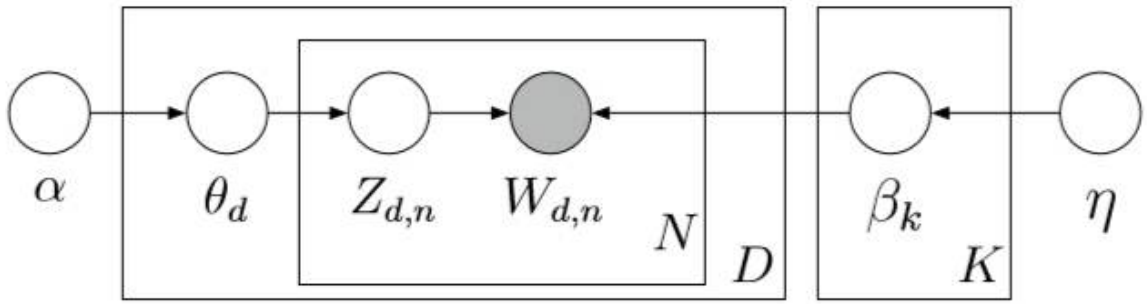


- Filtering



Topic Models





$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right)$$

Topic Models

NIPS 2012 papers
(in nicer format than this)
maintained by @karpathy
source code on github

Below every paper are TOP 100 most-occuring words in that paper and their color is based on LDA topic model with k = 7.
(It looks like 0 = theory, 1 = reinforcement learning, 2 = graphical models, 3 = deep learning/vision, 4 = optimization, 5 = neuroscience, 6 = embeddings etc.)

Toggle LDA topics to sort by: TOPIC0 TOPIC1 TOPIC2 TOPIC3 TOPIC4 TOPIC5 TOPIC6

Discriminatively Trained Sparse Code Gradients for Contour Detection
Ren Xiaofeng, Liefeng Bo

[pdf] [bibtex] [supplementary]
[rank by tf-idf similarity to this]
[abstract]

[set, algorithm, including] [average, approach, benchmark, evaluation] [comparing, normal, hierarchical] [contour, gpb, local, detection, depth, scg, color, image, oriented, matching, contrast, object, grayscale, precision, recognition, transform, work, learned, pooling, pixel, representation, double, global, learn, accuracy, scale, level, segmentation, figure, feature, nyu, globalization, scene, training, rich, single, automatically, apply, discriminative, codewords, ieee, half, directly, unsupervised, higher, chromaticity] [sparse, dictionary, gradient, pursuit, size, spectral, analysis, edge, step, sparsity] [power, coding, surface, natural] [code, learning, linear, data, orthogonal, dataset, svm, large, better, table, well, datasets]

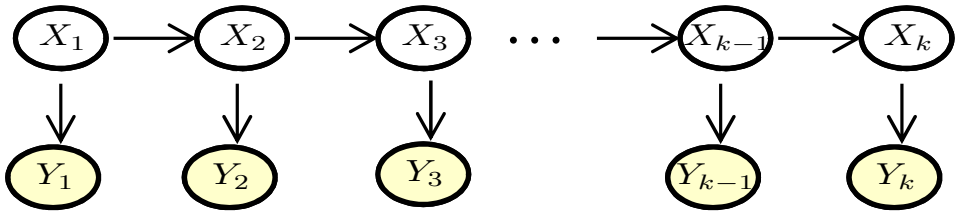
Deep Learning of Invariant Features via Simulated Fixations in Video
Will Zou, Andrew Ng, Shenghuo Zhu, Kai Yu

[pdf] [bibtex] [supplementary]
[rank by tf-idf similarity to this]
[abstract]

http://cs.stanford.edu/people/karpathy/nipspreview/

Filtering

- Hidden Markov Models (HMM) /Linear Dynamical Systems (LDS)



$$p(y_1 \ldots, y_K, x_1, \ldots, x_K) = p(x_1)p(y_1|x_1) \prod_{t=1}^{K-1} p(x_{t+1}|x_t)p(y_t|x_t)$$

HMM

$$p(x_1 = j) = \pi_j$$

$$p(x_{t+1}|x_t) = T_{ij}$$

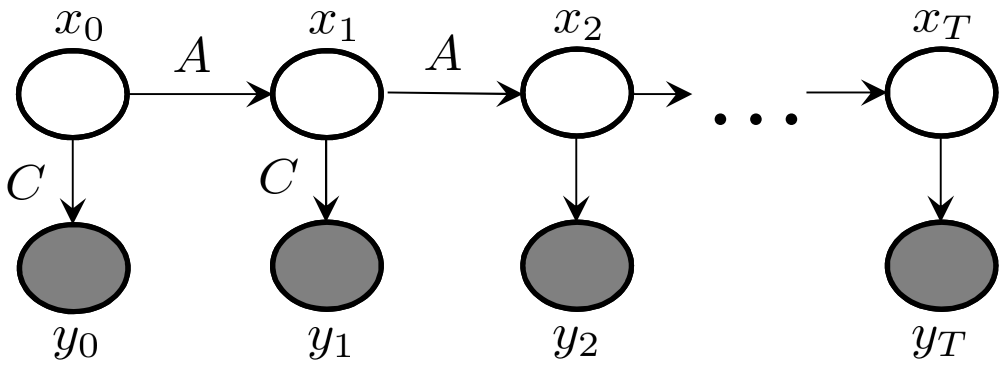
$$p(y_t|x_t) = A_j(y)$$

LDS

$$x_{t+1} = Ax_t + Gw_t \quad w_t \sim \mathcal{N}(0, Q)$$

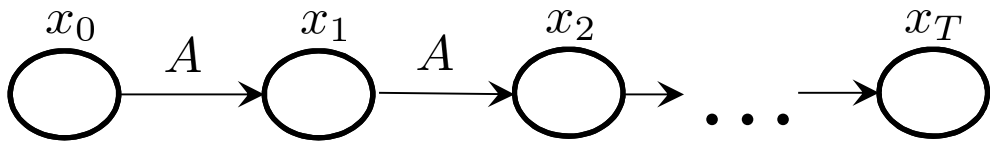
$$y_t = Cx_t + v_t \qquad v_t \sim \mathcal{N}(0, R)$$

Kalman Filter



$$x_{t+1} = Ax_t + Gw_t$$
$$w_t \sim \mathcal{N}(0, Q)$$
$$y_t = Cx_t + v_t$$
$$v_t \sim \mathcal{N}(0, R)$$

Unconstraied distribution



$$\mu_{t+1} = 0$$

$$\begin{aligned}\Sigma_{t+1} &= \mathbb{E}[x_{t+1}x_{t+1}^\top] = \mathbb{E}[(Ax_t + Gw_t)(Ax_t + Gw_t)^\top] \\ &= A\mathbb{E}[x_t x_t^\top]A^\top + G\mathbb{E}[w_t w_t^\top]G^\top \\ &= A\Sigma_t A^\top + GQG^\top\end{aligned}$$

Also, for joint density if necessary

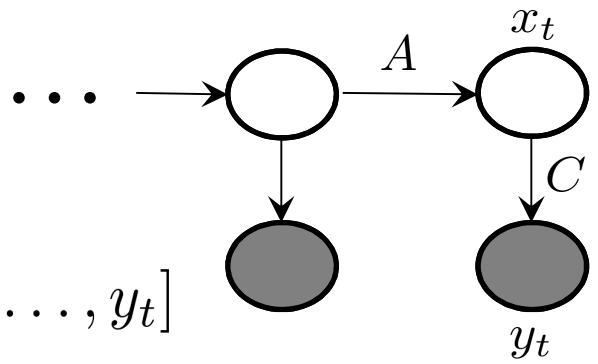
$$\mathbb{E}[x_t x_{t+1}^\top] = \Sigma_t A^\top$$

Kalman Filter

Filtering

$$\widehat{x}_{t|t} = \mathbb{E}[x_t | y_0, \ldots, y_t]$$

$$P_{t|t} = \mathbb{E}[(x_t - \widehat{x}_{t|t})(x_t - \widehat{x}_{t|t})^\top | y_0, \ldots, y_t]$$



"Conditional marginalization"
Marginalization from the left

$$\widehat{x}_{t|t} \;\&\; P_{t|t} \longrightarrow \widehat{x}_{t+1|t+1} \;\&\; P_{t+1|t+1}$$

Why filtering? Once we know $\hat{x}_{t|t} \;\&\; P_{t|t}$, we don't have to know (or keep) $y_0, \ldots, y_t$.

Kalman Filter

- Time update

$$p(x_t|y_0,\ldots,y_t) \rightarrow p(x_{t+1}|y_0,\ldots,y_t)$$

- Measurement update

$$p(x_{t+1}|y_0,\ldots,y_t) \rightarrow p(x_{t+1}|y_0,\ldots,y_t,y_{t+1})$$

Time update

$$\widehat{x}_{t+1|t} = A\widehat{x}_{t|t}$$

$$\begin{aligned} P_{t+1|t} &= \mathbb{E}[(x_{t+1}-\widehat{x}_{t+1|t})(x_{t+1}-\widehat{x}_{t+1|t})^\top|y_0,\ldots,y_t] \\ &= \mathbb{E}[(Ax_t+Gw_t-A\widehat{x}_{t|t})(Ax_t+Gw_t-A\widehat{x}_{t|t})^\top|y_0,\ldots,y_t] \\ &= AP_{t|t}A^\top+GQG^\top \end{aligned}$$

Kalman Filter

$$\mathbb{E}[y_{t+1}|y_0,\ldots,y_t] = \mathbb{E}[Cx_{t+1}+v_{t+1}|y_0,\ldots,y_t]$$
$$= C\widehat{x}_{t+1|t}$$

$$\mathbb{E}[(y_{t+1}-\widehat{y}_{t+1|t})(y_{t+1}-\widehat{y}_{t+1|t})^\top|y_0,\ldots,y_t]$$
$$= \mathbb{E}[(Cx_{t+1}+v_{t+1}-C\widehat{x}_{t+1|t})(Cx_{t+1}+v_{t+1}-C\widehat{x}_{t+1|t})^\top|y_0,\ldots,y_t]$$
$$= CP_{t+1|t}C^T + R$$

Also,

$$\mathbb{E}[(y_{t+1}-\widehat{y}_{t+1|t})(x_{t+1}-\widehat{x}_{t+1|t})^\top|y_0,\ldots,y_t]$$
$$= \mathbb{E}[(Cx_{t+1}+v_{t+1}-C\widehat{x}_{t+1|t})(x_{t+1}-\widehat{x}_{t+1|t})^\top|y_0,\ldots,y_t]$$
$$= CP_{t+1|t}$$

Joint:

$$p(x_{t+1},y_{t+1}|y_0,\ldots,y_t)$$
$$= \mathcal{N}\left(\begin{pmatrix}\widehat{x}_{t+1|t}\\ C\widehat{x}_{t+1|t}\end{pmatrix},\begin{pmatrix}P_{t+1|t} & P_{t+1|t}C^\top\\ CP_{t+1|t} & CP_{t+1|t}C^\top+R\end{pmatrix}\right)$$

Kalman Filter

Measurement update (Conditional density)

$$p(x_{t+1}|y_0,\ldots,y_{t+1}) = \mathcal{N}(\hat{x}_{t+1|t+1}, P_{t+1|t+1})$$

$$\begin{cases} \hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + P_{t+1|t}C^\top(CP_{t+1|t}C^\top + R)^{-1}(y_{t+1} - C\hat{x}_{t+1|t}) \\ P_{t+1|t+1} = P_{t+1|t} - P_{t+1|t}C^\top(CP_{t+1|t}C^\top + R)^{-1}CP_{t+1|t} \end{cases}$$

- Sum - ups

$$\hat{x}_{t+1|t} = A\hat{x}_{t|t}$$

$$P_{t+1|t} = AP_{t|t}A^\top + GQG^\top$$

$$\hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + P_{t+1|t}C^\top(CP_{t+1|t}C^\top + R)^{-1}(y_{t+1} - C\hat{x}_{t+1|t})$$

$$P_{t+1|t+1} = P_{t+1|t} - P_{t+1|t}C^\top(CP_{t+1|t}C^\top + R)^{-1}CP_{t+1|t}$$

Kalman Filter

- With different notation,

$$K_{t+1} \equiv P_{t+1|t} C^\top (C P_{t+1|t} C^\top + R)^{-1}$$

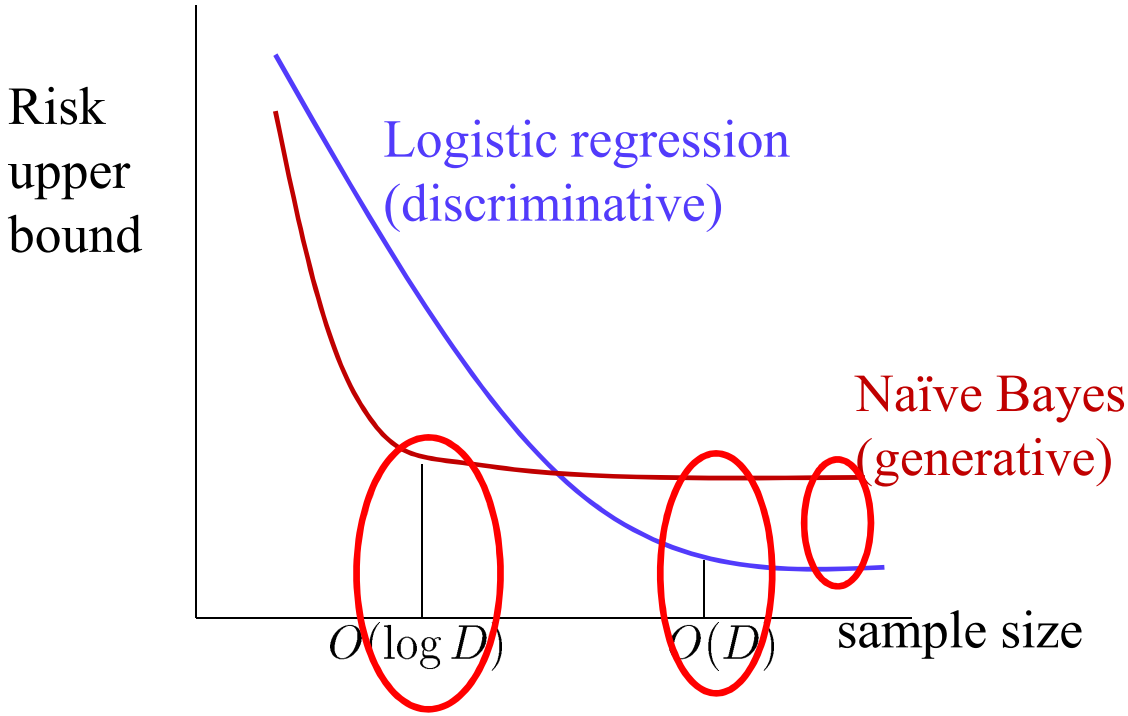
$$\widehat{x}_{t+1|t+1} = \widehat{x}_{t+1|t} + K_{t+1}(y_{t+1} - C\widehat{x}_{t+1|t})$$

- Alternative form of K_{t+1}

$$\begin{aligned} K_{t+1} &= P_{t+1|t} C^\top (C P_{t+1|t} C^\top + R)^{-1} \\ &= (P_{t+1|t}^{-1} + C^\top R C)^{-1} C^\top R^{-1} \\ &= (P_{t+1|t} + P_{t+1|t} C^\top (C P_{t+1|t} C^\top + R)^{-1} C P_{t+1|t}) C^\top R^{-1} \\ &= P_{t+1|t+1} C^\top R^{-1} \end{aligned}$$

Generative Model vs. Discriminative Model

- Comparative study of generative & discriminative pair
  - Same number of parameters, same form of *f(x)*



ANY QUESTIONS?