STA 414/2104: Machine Learning

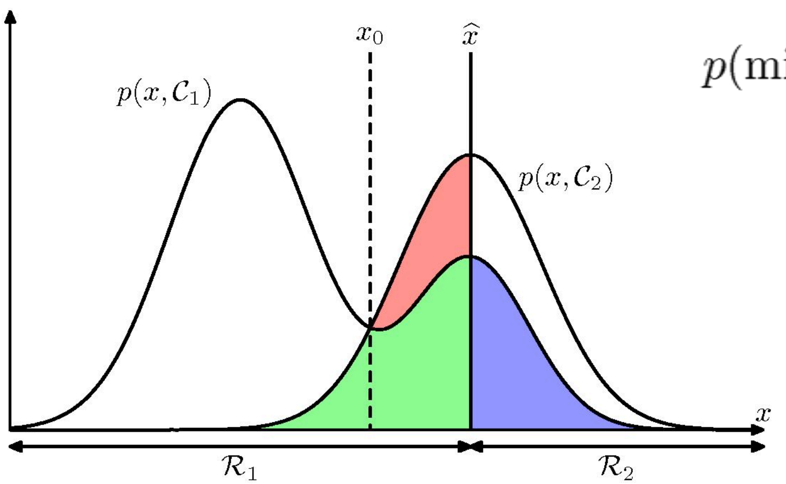
Russ Salakhutdinov

Department of Computer Science
Department of Statistics
rsalakhu@cs.toronto.edu
http://www.cs.toronto.edu/~rsalakhu/

Lecture 10

Final Review

- Polynomial curve fitting – generalization, overfitting
- Decision theory:
    - Minimizing misclassification rate / Minimizing the expected loss



$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}. \end{aligned}$$

- Loss functions for regression

$$\mathbb{E}[L] = \int\int \left(t - y(\mathbf{x})\right)^2 p(\mathbf{x}, t)\mathrm{d}\mathbf{x}\mathrm{d}t.$$

Final Review

- Bernoulli, Multinomial random variables (mean, variances)
- Multivariate Gaussian distribution (form, mean, covariance)
- Maximum likelihood estimation for these distributions.

- Exponential family / Maximum likelihood estimation / sufficient statistics for exponential family.

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

- Linear basis function models / maximum likelihood and least squares:

$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \sum_{i=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n), \beta)$$

$$= -\frac{\beta}{2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n)\right)^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi).$$

$$\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$

Final Review

Ridge regression

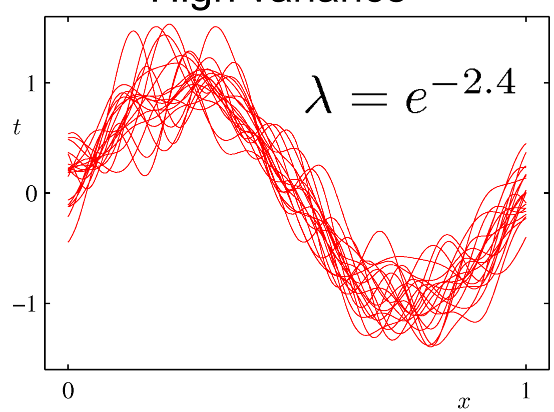
- Regularized least squares:

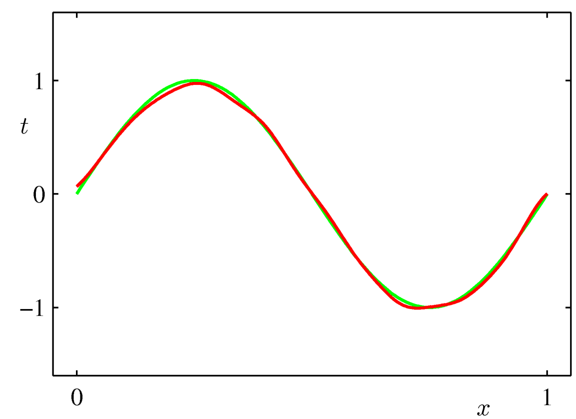
$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \qquad \mathbf{w} = \left(\lambda\mathbf{I} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}.$$

- Bayesian interpretation

- Bias-variance decomposition.

High variance





Low bias

Final Review

• Bayesian Inference: likelihood, prior, posterior:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})P(\mathbf{w})}{P(\mathcal{D})}$$

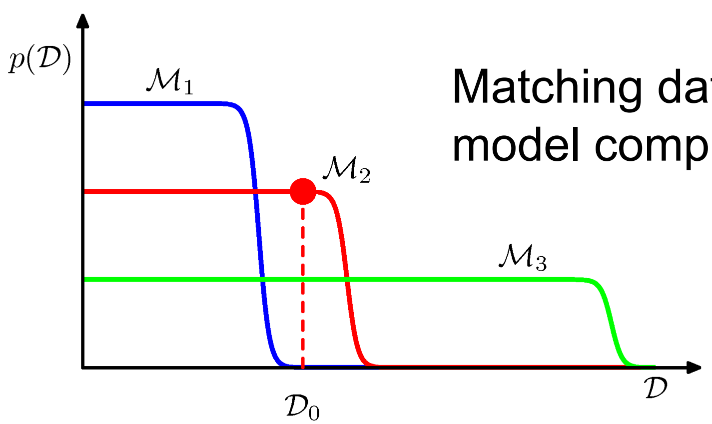
Marginal likelihood (normalizing constant):

$$P(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})P(\mathbf{w})\mathrm{d}\mathbf{w}$$

• Marginal likelihood / predictive distribution.

• Bayesian linear regression / parameter estimation / posterior distribution / predictive distribution

• Bayesian model comparison / Evidence approximation



Matching data and model complexity

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\mathbf{w}_{\mathrm{MAP}}) + M \ln \left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}} \right).$$

Final Review

- Classification models:
  - Discriminant functions
  - Fisher's linear discriminant

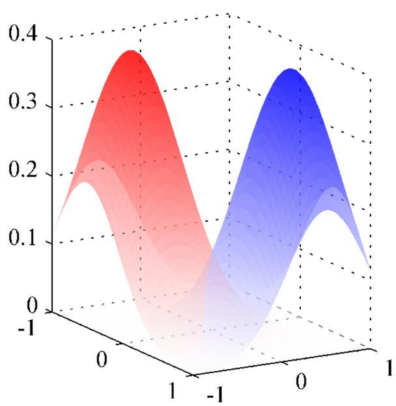
- Probabilistic Generative Models / Gaussian class conditionals / Maximum likelihood estimation:

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right).$$

$$p(\mathcal{C}_k|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + w_0),$$

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}.$$



Final Review

- Discriminative Models / Logistic regression / maximum likelihood estimation

Final Review

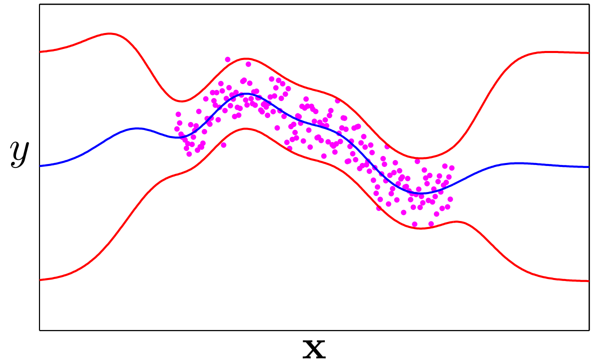
- Gaussian processes, definition:

$$\begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_N) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_N) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix} \right)$$

- GPs for regression.

- Marginal/predictive distributions. Making predictions using GPs.

- Covariance functions, automatic relevance determination, role of hyperparameters



$$p(f|\mathcal{D}) = \frac{p(f)p(\mathcal{D}|f)}{p(\mathcal{D})}$$

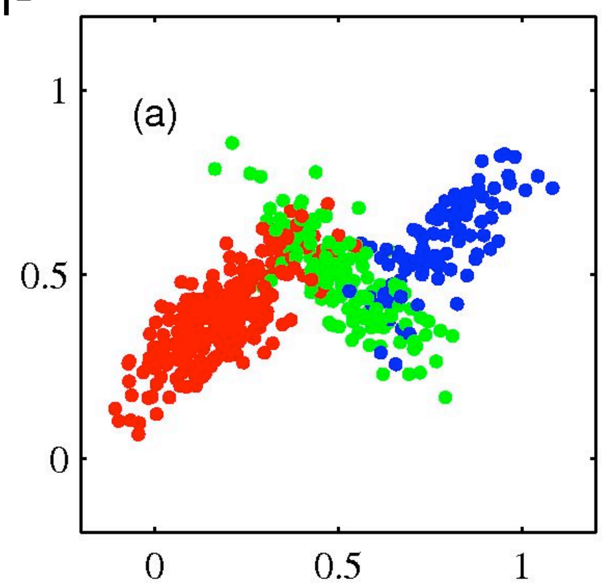
Final Review

• Mixture Models, k-means, Mixture of Gaussians

• Mixture of Gaussians: Maximum likelihood estimation.

• EM algorithm: definition of E-step, definition of M-step, relationship to k-means.

• Alternative view of EM: expected complete data log-likelihood:



• E-step: Compute posterior over latent variables: p(Z|X,θ^{old}).

• M-step: Find the new estimate of parameters θ^{new}:

$$\theta^{new} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{old}).$$

where

$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\theta).$$

Final Review

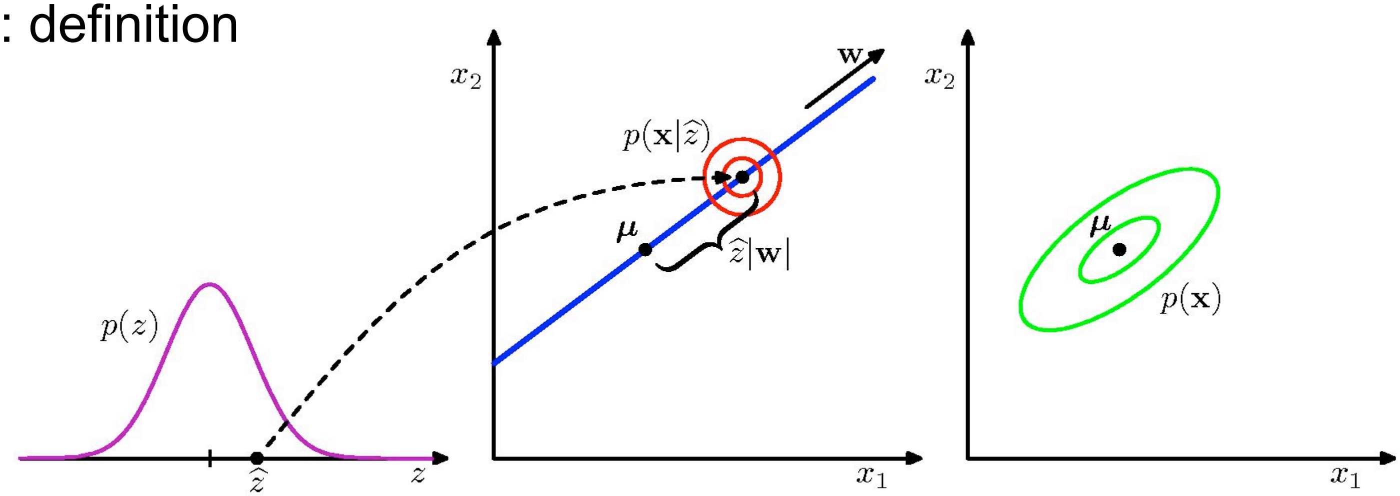
- Continuous latent variable models: Probabilistic PCA, Factor Analysis

- PCA, PCA for high-dimensional data
- Probabilistic PCA: definition of probabilistic model, Joint/Marginal density, posterior over latent variables, relationship to standard PCA, EM for PPCA.

- Probabilistic PCA: Maximum likelihood estimation, zero noise limit.

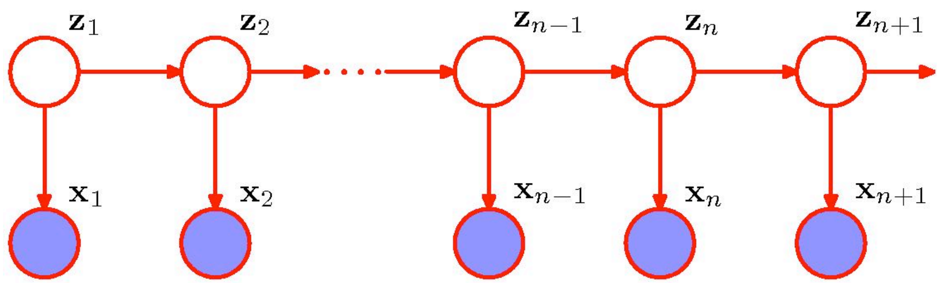
- Factor analysis, definition, marginal/joint/posterior. Relationship to PPCA.

- Autoencoders: definition



Final Review

- Sequential data: Markov models, maximum likelihood estimation
- State Space models: definition, transition model, observation model.



- Hidden Markov models: definition, transition model, observation model.
- Maximum likelihood estimation for HMMs, basics of EM algorithm.
- Basics of EM algorithm for HMMs: interring posterior over latent paths and parameter estimation for the transition and observation model.
- Dynamic programming (understanding of alpha-beta recursions)
- Viterbi decoding.