Machine Learning

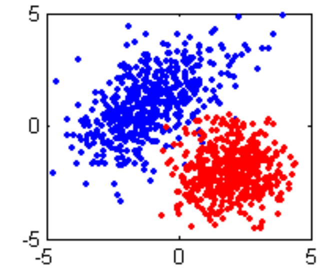
4771

Instructor: Tony Jebara

Topic 10

- Classification with Gaussians
- Regression with Gaussians
- Principal Components Analysis

Classification with Gaussians



- Have two classes, each with their own Gaussian:

$$\left\{\left(x_1,y_1\right),\ldots,\left(x_N,y_N\right)\right\} \quad x\in R^D \;\; y\in\left\{0,1\right\}$$

- Given parameters $\theta=\left\{\alpha,\mu_0,\Sigma_0,\mu_1,\Sigma_1\right\}$ we can generate iid data from $p\left(x,y\mid\theta\right)=p\left(y\mid\theta\right)p\left(x\mid y,\theta\right)$ by:

  1) flipping a coin to get y via Bernoulli $p\left(y\mid\theta\right)=\alpha^y\left(1-\alpha\right)^{1-y}$

  2) sampling an x from y'th Gaussian $p\left(x\mid y,\theta\right)=N\left(x\mid\mu_y,\Sigma_y\right)$

- Or, recover parameters from data using maximum likelihood

$$l\left(\theta\right)=\log p\left(data\mid\theta\right)=\sum_{i=1}^N\log p\left(x_i,y_i\mid\theta\right)$$
$$=\sum_{i=1}^N\log p\left(y_i\mid\theta\right)+\sum_{i=1}^N\log p\left(x_i\mid y_i,\theta\right)$$
$$=\sum_{i=1}^N\log p\left(y_i\mid\alpha\right)+\sum_{y_i\in 0}\log p\left(x_i\mid\mu_0,\Sigma_0\right)+\sum_{y_i\in 1}\log p\left(x_i\mid\mu_1,\Sigma_1\right)$$

Classification with Gaussians

- Max Likelihood can be done separately for the 3 terms

$$l = \sum_{i=1}^{N} \log p\left(y_i \mid \alpha\right) + \sum_{y_i \in 0} \log p\left(x_i \mid \mu_0, \Sigma_0\right) + \sum_{y_i \in 1} \log p\left(x_i \mid \mu_1, \Sigma_1\right)$$

- Count # of pos & neg examples (class prior): $\alpha = \frac{N_1}{N_0 + N_1}$
- Get mean & cov of negatives and mean & cov of positives:

$$\mu_0 = \frac{1}{N_0} \sum_{y_{i \in 0}} x_i \qquad \Sigma_0 = \frac{1}{N_0} \sum_{y_{i \in 0}} \left(x_i - \mu_0\right)\left(x_i - \mu_0\right)^T$$

$$\mu_1 = \frac{1}{N_1} \sum_{y_{i \in 1}} x_i \qquad \Sigma_1 = \frac{1}{N_1} \sum_{y_{i \in 1}} \left(x_i - \mu_1\right)\left(x_i - \mu_1\right)^T$$

- Given (x,y) pair, can now compute likelihood $p\left(x, y\right)$
- To make classification, a bit of Decision Theory
- Without x, can compute prior guess for y $p\left(y\right)$
- Give me x, want y, I need posterior $p\left(y \mid x\right)$
- Bayes Optimal Decision: $\hat{y} = \arg\max_{y = \{0,1\}} p\left(y \mid x\right)$
- Optimal iff we have true probability

Posterior gives Logistic

- Bayes Optimal Decision: $\hat{y} = \arg\max_{y=\{0,1\}} p(y \mid x)$
- To get conditional:

$$p(y \mid x) = \frac{p(x,y)}{p(x)} = \frac{p(x,y)}{\sum_y p(x,y)} = \frac{p(x,y)}{p(x,y=0) + p(x,y=1)}$$

- Check which is greater: $p(y=0 \mid x) \geq ? \leq p(y=1 \mid x)$
- Or check if this is > 0.5

$$p(y=1 \mid x) = \frac{p(x,y=1)}{p(x,y=0) + p(x,y=1)}$$

$$= \frac{1}{\frac{p(x,y=0)}{p(x,y=1)} + 1}$$

$$= \frac{1}{\exp\left(-\log \frac{p(x,y=1)}{p(x,y=0)}\right) + 1}$$

- Get logistic squashing function of log-ratio of probability models

$$= sigmoid\left(\log \frac{p(x,y=1)}{p(x,y=0)}\right)$$

Linear or Quadratic Decisions

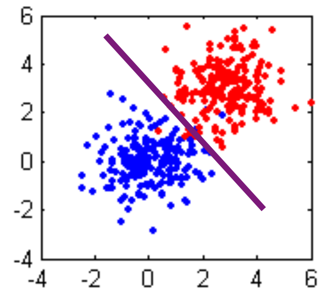
- Example cases, plotting decision boundary when = 0.5

$$p\left(y=1 \mid x\right) = \frac{p\left(x,y=1\right)}{p\left(x,y=0\right)+p\left(x,y=1\right)}$$

$$= \frac{\alpha N\left(x \mid \mu_1,\Sigma_1\right)}{\left(1-\alpha\right)N\left(x \mid \mu_0,\Sigma_0\right)+\alpha N\left(x \mid \mu_1,\Sigma_1\right)}$$

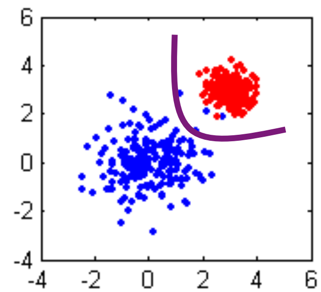
- If covariances are equal:

  linear decision



- If covariances are different:

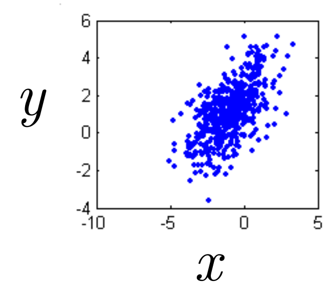
  quadratic decision



Regression with Gaussians

- Have input and output, each Gaussian:

$$\left\{\left(x_1,y_1\right),\ldots,\left(x_N,y_N\right)\right\} \quad x \in R^{D_x} \quad y \in R^{D_y}$$



concatenate $z_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}$

$$p\left(z \mid \mu,\Sigma\right) = \frac{1}{\left(2\pi\right)^{D/2}\sqrt{\left|\Sigma\right|}} \exp\left(-\tfrac{1}{2}\left(z-\mu\right)^T \Sigma^{-1}\left(z-\mu\right)\right)$$

- Maximum Likelihood is as usual for a multivariate Gaussian

$$\mu = \tfrac{1}{N}\sum_{i=1}^{N} z_i \qquad \Sigma = \tfrac{1}{N}\sum_{i=1}^{N}\left(z_i-\mu\right)\left(z_i-\mu\right)^T$$

- Bayes optimal decision: $\hat{y} = \arg\max_{y\in\mathbb{R}} p\left(y \mid x\right)$
- Or we can use: $\hat{y} = E_{p(y|x)}\left\{y\right\}$
- Have joint, need conditional: $p\left(y \mid x\right) = \dfrac{p\left(x,y\right)}{p\left(x\right)} = \dfrac{p\left(x,y\right)}{\int_y p\left(x,y\right)}$

Gaussian Marginals/Conditionals

- Conditional & marginal from joint: $p\left(y \mid x\right) = \frac{p\left(x,y\right)}{p\left(x\right)} = \frac{p\left(x,y\right)}{\int_y p\left(x,y\right)}$

- Conditioning the Gaussian:

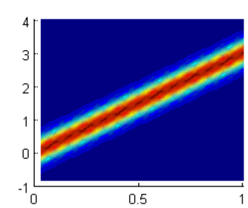
$$p\left(z \mid \mu, \Sigma\right) = \frac{1}{\left(2\pi\right)^{D/2}\sqrt{\left|\Sigma\right|}} \exp\left(-\frac{1}{2}\left(z-\mu\right)^T \Sigma^{-1}\left(z-\mu\right)\right)$$

$$p\left(x,y\right) = \frac{1}{\left(2\pi\right)^{D/2}\sqrt{\left|\Sigma\right|}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}\right)^T \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}^{-1} \left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}\right)\right)$$

$$p\left(x\right) = \frac{1}{\left(2\pi\right)^{D_x/2}\sqrt{\left|\Sigma_{xx}\right|}} \exp\left(-\frac{1}{2}\left(x-\mu_x\right)^T \Sigma_{xx}^{-1}\left(x-\mu_x\right)\right)$$

$$= N\left(x \mid \mu_x, \Sigma_{xx}\right)$$

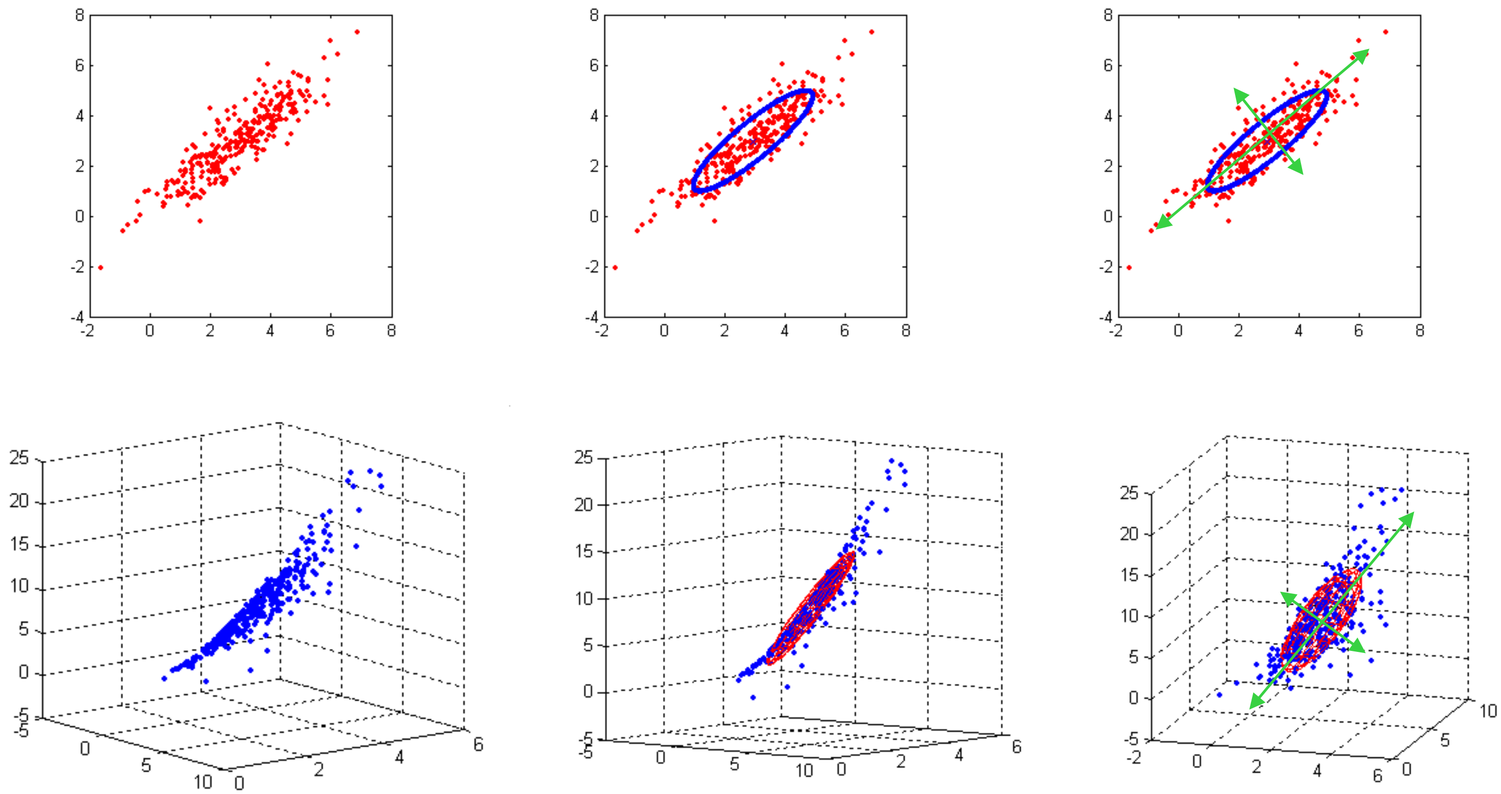
$$p\left(y \mid x\right) = N\left(y \mid \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}\left(x-\mu_x\right), \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right)$$



- Here argmax is expectation which is conditional mean: $\hat{y} = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}\left(x-\mu_x\right)$

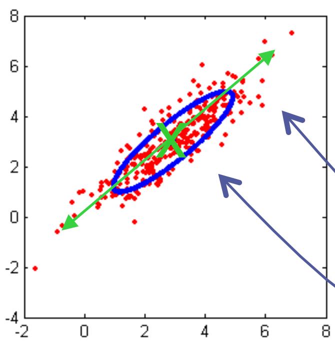
Principal Components Analysis

- Gaussians: for Classification, Regression… & Compression!
- Data can be constant in some directions, changes in others
- Use Gaussian to find directions of high/low variance

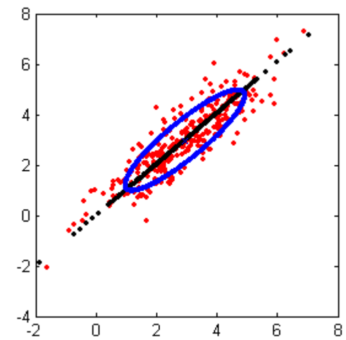


Principal Components Analysis

- Idea: instead of writing data in all its dimensions, only write it as mean + steps along one direction

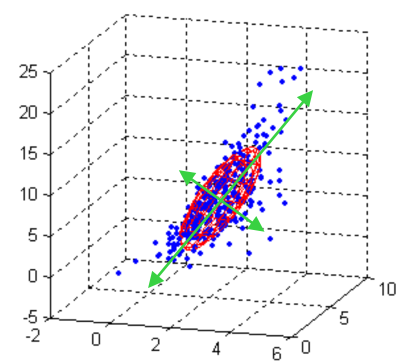


$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} \approx \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} + c_i \begin{bmatrix} v_x \\ v_y \end{bmatrix}$$



- More generally, keep a subset of dimensions C from D (i.e. 2 of 3)

$$\vec{x}_i \approx \vec{\mu} + \sum_{j=1}^{C} c_{ij} \vec{v}_j$$



- Compression method: $\vec{x}_i \gg \vec{c}_i$
- Optimal directions: along eigenvectors of covariance
- Which directions to keep: highest eigenvalues (variances)

Principal Components Analysis

- If we have eigenvectors, mean and coefficients:

$$\vec{x}_i \approx \vec{\mu} + \sum_{j=1}^{C} c_{ij} \vec{v}_j$$

- Get eigenvectors (use eig() in Matlab): $\Sigma = V \Lambda V^T$

$$\begin{bmatrix} \Sigma(1,1) & \Sigma(1,2) & \Sigma(1,3) \\ \Sigma(1,2) & \Sigma(2,2) & \Sigma(2,3) \\ \Sigma(1,3) & \Sigma(2,3) & \Sigma(3,3) \end{bmatrix} = \begin{bmatrix} [\vec{v}_1] & [\vec{v}_2] & [\vec{v}_3] \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \begin{bmatrix} [\vec{v}_1] & [\vec{v}_2] & [\vec{v}_3] \end{bmatrix}^T$$

- Eigenvectors are orthonormal: $\vec{v}_i^T \vec{v}_j = \delta_{ij}$
- In coordinates of v, Gaussian is diagonal, cov = Λ
- All eigenvalues are non-negative $\lambda_i \geq 0$
- Higher eigenvalues are higher variance, use the top C ones

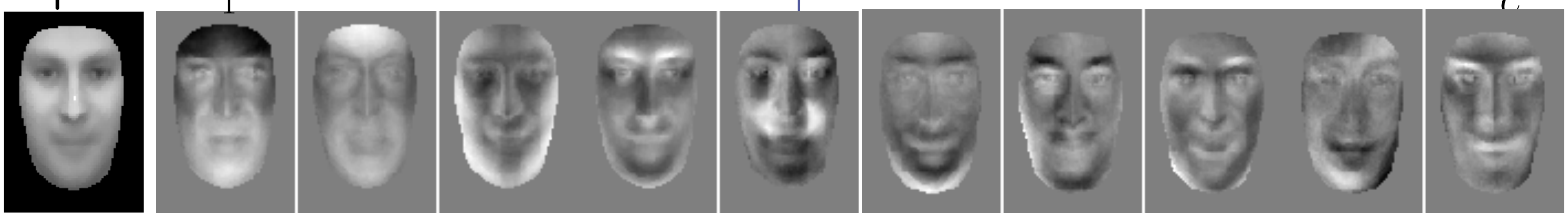
$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4 \geq \ldots$$

- To compute the coefficients: $c_{ij} = \left(\vec{x}_i - \vec{\mu}\right)^T \vec{v}_j$

Eigenfaces

$$\left\{x_1, \ldots, x_N\right\}$$

$\vec{\mu}$ $\vec{v}_1$ $\vec{v}_C$



ENCODE

$$\boxed{c_{ij} = \left(\vec{x}_i - \vec{\mu}\right)^T \vec{v}_j}$$

DECODE

$$\left\{\left(\hat{x}_1 = \mu + \sum_{j=1}^{C} c_{1j}\vec{v}_j\right), \ldots, \left(\hat{x}_N = \mu + \sum_{j=1}^{C} c_{Nj}\vec{v}_j\right)\right\}$$

Machine Learning

4771

Instructor: Tony Jebara

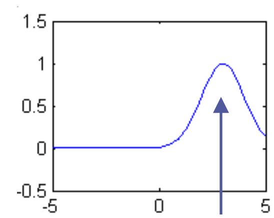
Topic 11

- Maximum Likelihood as Bayesian Inference
- Maximum A Posteriori
- Bayesian Gaussian Estimation

Why Maximum Likelihood?

- So far, assumed max (log) likelihood (IID or otherwise)
- Philosophical: Why?

$$\max_\theta L(\theta) = \max_\theta p(x_1,\ldots,x_N \mid \theta)$$
$$= \max_\theta \prod_{i=1}^{N} p(x_i \mid \theta)$$



- Also, why ignore p(θ)?

- Hint: Recall Bayes rule:

$$\underbrace{p(\theta \mid x)}_{\text{posterior}} = \frac{\overbrace{p(x \mid \theta)}^{\text{likelihood}}\,\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(x)}_{\text{evidence}}}$$

- Everyone agrees on probability theory: inference and use of probability models when we have computed p(x)
- But how get to p(x) from data? Debate…
- Two schools of thought: Bayesians and Frequentists

Bayesians & Frequentists

•Frequentists (Neymann/Pearson/Wald). An orthodox view that sampling is infinite and decision rules can be sharp.

•Bayesians (Bayes/Laplace/de Finetti). Unknown quantities are treated probabilistically and the state of the world can always be updated.

de Finetti: p(event) = price I would pay for a contract that pays 1$ when event happens

•Likelihoodists (Fisher). Single sample inference based on maximizing the likelihood function and relying on the Birnbaum's Theorem. Bayesians – But they don't know it.

Bayesians & Frequentists

- Frequentists:
  - Data are a repeatable random sample- there is a frequency
  - Underlying parameters remain constant during this repeatable process
  - Parameters are fixed

- Bayesians:
  - Data are observed from the realized sample.
  - Parameters are unknown and described probabilistically
  - Data are fixed

Bayesians & Frequentists

- Frequentists: classical / objective view / no priors
  - every statistician should compute same p(x) so no priors
  - can't have a p(event) if it never happened
  - avoid $p(\theta)$, there is 1 true model, not distribution of them
  - permitted: $p_\theta(x,y)$ forbidden: $p(x,y|\theta)$
  - Frequentist inference: estimate one best model θ
    - use the ML estimator (unbiased & minimum variance)
    - do not depend on Bayes rule for learning
- Bayesians: subjective view / priors are ok
  - put a distribution or pdf on all variables in the problem
  - even models & deterministic quantities (i.e. speed of light)
  - use a prior $p(\theta)$, on the model θ before seeing any data
  - Bayesian inference: use Bayes rule for learning, integrate over all model (θ) unknown variables

Bayesian Inference

- Bayes rule gives rise to maximum likelihood
- Assume we have a prior over models $p(\theta)$

likelihood

$$p\left(\theta \mid x\right) = \frac{p\left(x \mid \theta\right) p\left(\theta\right)}{p\left(x\right)}$$

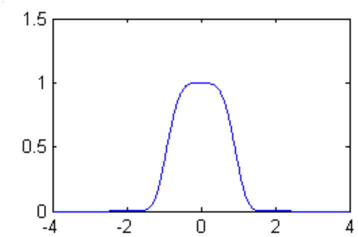
prior

posterior

evidence

- How to pick $p(\theta)$?
  - Pick simpler θ is better
  - Pick form for mathematical convenience

$p\left(\theta\right)$



- We have data (can assume IID): $\mathfrak{X} = \left\{x_1, x_2, \ldots, x_N\right\}$
- Want to get a model to compute: $p\left(x\right)$
- Want p(x) given our data… How to proceed?

Bayesian Inference

- Want p(x) given our data… $p\left(x \mid \mathcal{X}\right) = p\left(x \mid x_1, x_2, \ldots, x_n\right)$

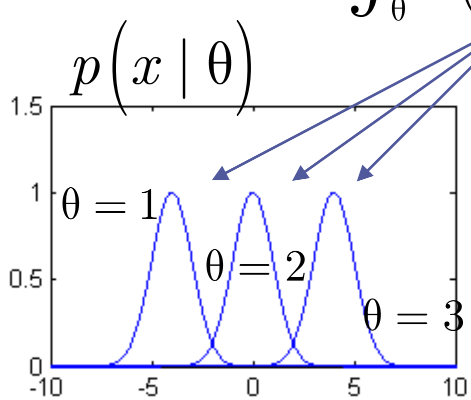
$$p\left(x \mid \mathcal{X}\right) = \int_\theta p\left(x, \theta \mid \mathcal{X}\right) d\theta$$

$$= \int_\theta p\left(x \mid \theta, \mathcal{X}\right) p\left(\theta \mid \mathcal{X}\right) d\theta$$

Prior

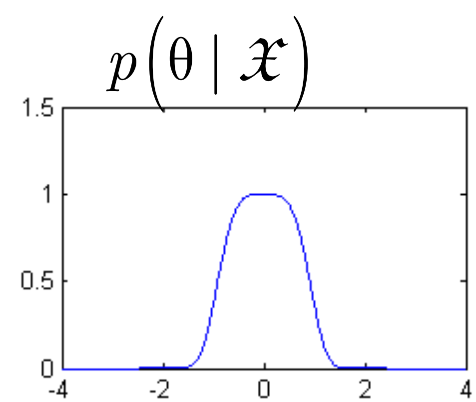
$$= \int_\theta p\left(x \mid \theta, \mathcal{X}\right) \frac{p\left(\mathcal{X} \mid \theta\right) p\left(\theta\right)}{p\left(\mathcal{X}\right)} d\theta$$

$$= \int_\theta p\left(x \mid \theta\right) \frac{\prod_{i=1}^{N} p\left(x_i \mid \theta\right) p\left(\theta\right)}{p\left(\mathcal{X}\right)} d\theta$$



Many models

Weight on each model



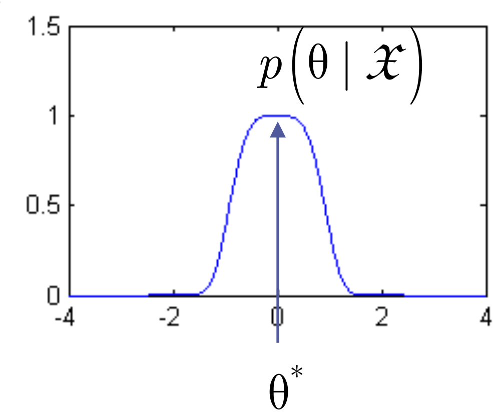
Bayesian Inference to MAP & ML

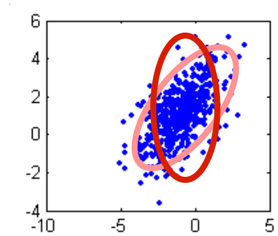
- The full Bayesian Inference integral can be mathematically tricky. Maximum likelihood is an approximation of it…

$$p\left(x \mid \mathcal{X}\right) = \int_\theta p\left(x \mid \theta\right) \frac{\prod_{i=1}^{N} p\left(x_i \mid \theta\right) p\left(\theta\right)}{p\left(\mathcal{X}\right)} d\theta$$

$$\approx \int_\theta p\left(x \mid \theta\right) \delta\left(\theta - \theta^*\right) d\theta$$

$$where\ \theta^* = \begin{cases} \arg\max_\theta \dfrac{\prod_{i=1}^{N} p\left(x_i \mid \theta\right) p\left(\theta\right)}{p\left(\mathcal{X}\right)} & MAP \\ \arg\max_\theta \dfrac{\prod_{i=1}^{N} p\left(x_i \mid \theta\right) uniform\left(\theta\right)}{p\left(\mathcal{X}\right)} & ML \end{cases}$$





- Maximum A Posteriori (MAP) is like Maximum Likelihood (ML) with a prior p(θ) which lets us prefer some models over others

$$l_{MAP}\left(\theta\right) = l_{ML}\left(\theta\right) + \log p\left(\theta\right) = \sum_{i=1}^{N} \log p\left(x_i \mid \theta\right) + \log p\left(\theta\right)$$

Bayesian Inference Example

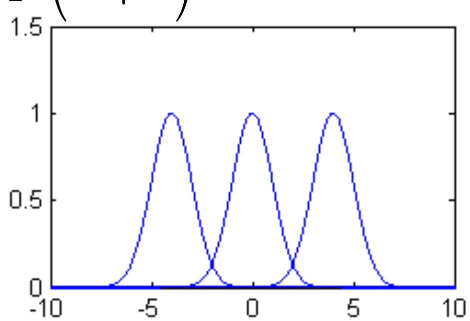
- For Gaussians, we CAN compute the integral (but hard!)

$$p(x \mid \mathcal{X}) = \int_\theta p(x \mid \theta) \frac{\prod_{i=1}^{N} p(x_i \mid \theta) p(\theta)}{p(\mathcal{X})} d\theta$$

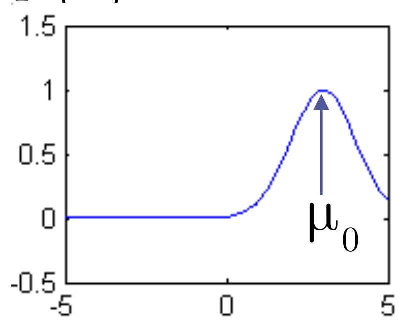
$$\propto \int_\theta p(x \mid \theta) \prod_{i=1}^{N} p(x_i \mid \theta) p(\theta) d\theta$$

- Example:... assume 1d Gaussian & Gaussian prior on mean

$p(x \mid \theta) = Gaussian$



$p(\theta) = Gaussian$



$$p(x \mid \mathcal{X}) \propto \int_\mu \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2} \right) \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\mu)^2} \right) \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\mu-\mu_0)^2} \right) d\mu$$

Tony Jebara, Columbia University

Bayesian Inference Example

- Solve integral over all Gaussian means with variance=1

$$p\left(x \mid \mathcal{X}\right) \propto \int_{\mu=-\infty}^{\mu=\infty} \left(\tfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(x-\mu\right)^2}\right) \prod_{i=1}^{N} \left(\tfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(x_i-\mu\right)^2}\right) \left(\tfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\mu_0-\mu\right)^2}\right) d\mu$$

$$\propto \int_{\mu=-\infty}^{\mu=\infty} \exp\left(-\tfrac{1}{2}\left(x-\mu\right)^2 - \sum_i \tfrac{1}{2}\left(x_i-\mu\right)^2 - \tfrac{1}{2}\left(\mu_0-\mu\right)^2\right) d\mu$$

$$\propto \int_{\mu=-\infty}^{\mu=\infty} \exp\left(-\tfrac{1}{2}\left[\left(N+2\right)\mu^2 - 2\mu\left(x+\mu_0+\sum_i x_i\right) + x^2\right]\right) d\mu$$

$$\propto \int_{\mu=-\infty}^{\mu=\infty} \exp\left(-\tfrac{1}{2}\left[\left(N+2\right)\mu^2 - 2\mu\left(x+\mu_0+\sum_i x_i\right) + x^2\right] + \left[\quad\right]^2 - \left[\quad\right]^2\right) d\mu$$

$$\propto \exp\left(-\tfrac{1}{2}\left[\frac{-\left(x+\mu_0+\sum_i x_i\right)^2}{N+2} + x^2\right]\right) \qquad \tilde{\mu} = \frac{\mu_0 + \sum_i x_i}{N+1}$$

$$= N\left(x \mid \tilde{\mu}, \tilde{\sigma}^2\right) \qquad \tilde{\sigma}^2 = \frac{N+2}{N+1}$$

- Can integrate over μ and Σ for multivariate Gaussian (Jordan ch. 4 and Minka Tutorial)

$$p\left(x \mid \mathcal{X}\right) = \frac{\Gamma\left(\left(N+1\right)/2\right)}{\Gamma\left(\left(N+1-d\right)/2\right)} \left|\frac{1}{\left(N+1\right)\pi} \bar{\Sigma}^{-1}\right|^{1/2} \left(\frac{1}{N+1}\left(x-\bar{\mu}\right)^T \bar{\Sigma}^{-1}\left(x-\bar{\mu}\right)+1\right)^{-\left(N+1\right)/2}$$