Lecture 6
Some Concepts Revisit

Rui Xia
School of Computer Science & Engineering
Nanjing University of Science & Technology
http://www.nustm.cn/~rxia

Outline

- Review of Linear Models
  - Linear Regression
  - Logistic Regression
  - Perceptron
- Generative vs. Discriminative
  - Hypothesis
  - Decision
  - Learning
- Over-fitting
  - ML - MAP
  - Regularization

3 Key Concepts in Machine Learning

- Hypothesis
  - Math models with (unknown) parameters (or structures)
- Learning (to estimate the parameters)
  - Maximum Likelihood Estimation (MLE), MAP, Bayesian Estimation
  - Cost Function Optimization
- Decision
  - Bayes decision rule
  - Direct prediction function

Model Hypothesis

- Linear Regression

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}$$

- Perceptron Algorithm

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \begin{cases} 1, & if\ \ \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x} \geq 0 \\ 0, & if\ \ \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x} < 0 \end{cases}$$

- Logistic Regression

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \delta\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}\right) = \frac{1}{1 + e^{-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}}}$$

$$P(y = 1|\boldsymbol{x};\boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(\boldsymbol{x}) \qquad P(y = 0|\boldsymbol{x};\boldsymbol{\theta}) = 1 - h_{\boldsymbol{\theta}}(\boldsymbol{x})$$

Learning Criteria (Cost Functions)

- Linear Regression

$$J_l(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{N}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2$$

Maximum Likelihood ⇔ Least Mean Square

Learning Criteria (Cost Functions)

- Perceptron Algorithm

$$
\begin{aligned}
J_p(\boldsymbol{\theta}) &= \sum_{\boldsymbol{x}^{(i)} \in M_0} \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}^{(i)} - \sum_{\boldsymbol{x}^{(j)} \in M_1} \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}^{(j)} \\
&= \sum_{i=1}^{N} \left(\left(1 - y^{(i)}\right) h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)} \left(1 - h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)\right) \right) \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}^{(i)} \\
&= \sum_{i=1}^{N} \left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right) \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}^{(i)}
\end{aligned}
$$

Perceptron Criterion

Learning Criteria (Cost Functions)

- Logistic Regression

$$J_c(\boldsymbol{\theta}) = \sum_{i=1}^{N} y^{(i)} \log h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)\right)$$

Maximum Likelihood ⇔ Minimum Cross Entropy Error

Gradient Descent Optimization

- Linear Regression

$$\frac{\partial}{\partial \boldsymbol{\theta}} J_l(\boldsymbol{\theta}) = \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{N} \left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2$$

$$= \sum_{i=1}^{N} \left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)\boldsymbol{x}^{(i)}$$

$$\boldsymbol{\theta} \coloneqq \boldsymbol{\theta} - \alpha \frac{\partial}{\partial \boldsymbol{\theta}} J_l(\boldsymbol{\theta}) = \boldsymbol{\theta} - \alpha \sum_{i=1}^{N} \left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)\boldsymbol{x}^{(i)}$$

(Stochastic) Gradient Descent Optimization

- Perceptron Algorithm

$$\frac{\partial}{\partial \boldsymbol{w}} J_p(\boldsymbol{w}) = \sum_{i=1}^{N} \left(h_{\boldsymbol{w}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)} \right) \boldsymbol{x}^{(i)}$$

$$\Downarrow$$

$$\boldsymbol{w} \coloneqq \boldsymbol{w} + \alpha\left(y - h_{\boldsymbol{w}}(\boldsymbol{x})\right)\boldsymbol{x}$$

$$\coloneqq \begin{cases} \boldsymbol{w} + \alpha \boldsymbol{x}, & if\ y = 1\ and\ h_{\boldsymbol{w}}(\boldsymbol{x}) = 0 \\ \boldsymbol{w} + \alpha \boldsymbol{x}, & if\ y = 0\ and\ h_{\boldsymbol{w}}(\boldsymbol{x}) = 1 \\ \boldsymbol{w}, & otherwise \end{cases}$$

Gradient Descent Optimization

- Logistic Regression

$$\frac{\partial J_c(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{N} \left(y^{(i)} \frac{1}{h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})} - \left(1 - y^{(i)}\right) \frac{1}{1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})} \right) \frac{\partial}{\partial \boldsymbol{\theta}} h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)$$

$$= \sum_{i=1}^{N} \left(y^{(i)} \frac{1}{h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})} - \left(1 - y^{(i)}\right) \frac{1}{1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})} \right) h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) \left(1 - h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)\right) \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}^{\mathrm{T}} x^{(i)}$$

$$= \sum_{i=1}^{N} \left(y^{(i)} \left(1 - h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)\right) - \left(1 - y^{(i)}\right) h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) \right) \boldsymbol{x}^{(i)}$$

$$= \sum_{i=1}^{N} \left(y - \boldsymbol{h}_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(\boldsymbol{i})}\right) \right) \boldsymbol{x}^{(i)}$$

$$\Downarrow$$

$$\boldsymbol{\theta} := \boldsymbol{\theta} + \alpha \sum_{i=1}^{N} \left(y^{(i)} - h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) \right) \boldsymbol{x}^{(i)}$$

Outline

- Review of Linear Models
  - Linear Regression
  - Logistic Regression
  - Perceptron
- Generative vs. Discriminative
  - Hypothesis
  - Decision
  - Learning
- Over-fitting
  - ML - MAP
  - Regularization

Hypothesis - Learning - Decision

- Discriminative Model
  - Directly Modeling Predictive Function

$$y = f(\boldsymbol{x})$$

Example:
Perceptron, SVMs

  - Modeling Conditional Distribution

$$p(y|\boldsymbol{x})$$

Example:
Logistic/Softmax Regression

- Generative Model (Modeling Joint Distribution)

$$p(\boldsymbol{x}, y) = p(y)p(\boldsymbol{x}|y)$$

Examples:
Naïve Bayes, GMM

Hypothesis - Learning - Decision

- Discriminative Model
  - Modeling Predictive Function

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

Optimizing some loss functions, such as least mean square (LMS), cross entropy (CE), Maximum Margin, etc.

  - Modeling Posterior Distribution

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_i \log p(y^{(i)}|\boldsymbol{x}^{(i)})$$

ML, MAP (for posterior distribution) ⇔ Equivalent to some criteria in some cases

- Generative Model (Modeling Joint/Marginal Distribution)

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_i \log p(\boldsymbol{x}^{(i)}, y^{(i)})$$

ML, MAP, Bayesian Inference (for joint or marginal distribution)

Hypothesis - Learning - Decision

- Discriminative Model
  - Conditional Distribution

$$\arg\max_{y} p(y|\boldsymbol{x})$$

  - Predictive Function

$$y = f(\boldsymbol{x})$$

- Generative Model
  - Bayes Formula

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})}$$

$$\Downarrow$$

$$\arg\max_{y} p(y|\boldsymbol{x}) = \arg\max_{y} p(\boldsymbol{x}, y) = \arg\max_{y} p(\boldsymbol{x}|y)p(y)$$

Generative Models for Classification

- Modeling Joint Distribution

Class-conditional probability

$$p(\boldsymbol{x}, y = c_j) = \boxed{p(c_j)}\boxed{p(\boldsymbol{x}|c_j)}$$

Class prior probability

- Different Class-Conditional Distribution
  - Multinomial Distribution

$$\begin{aligned} p(\boldsymbol{x}, c_j|\boldsymbol{\theta}) &= p(c_j|\boldsymbol{\theta})p(\boldsymbol{x}|c_j;\boldsymbol{\theta}) \\ &= p_j \prod_{t=1}^{M} \theta_{t,j}^{N(w_t,\boldsymbol{x})} \end{aligned}$$

  - Gaussian Distribution

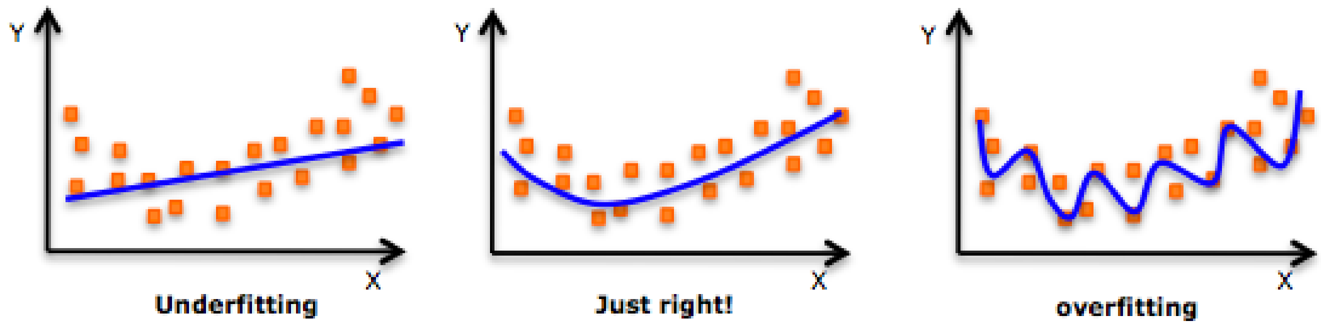
$$\begin{aligned} p(\boldsymbol{x}, c_j|\boldsymbol{\theta}) &= p(c_j|\boldsymbol{\theta})p(\boldsymbol{x}|c_j;\boldsymbol{\theta}) \\ &= p_j N(\boldsymbol{x}|\mu_j, \Sigma_j) \end{aligned}$$

Outline

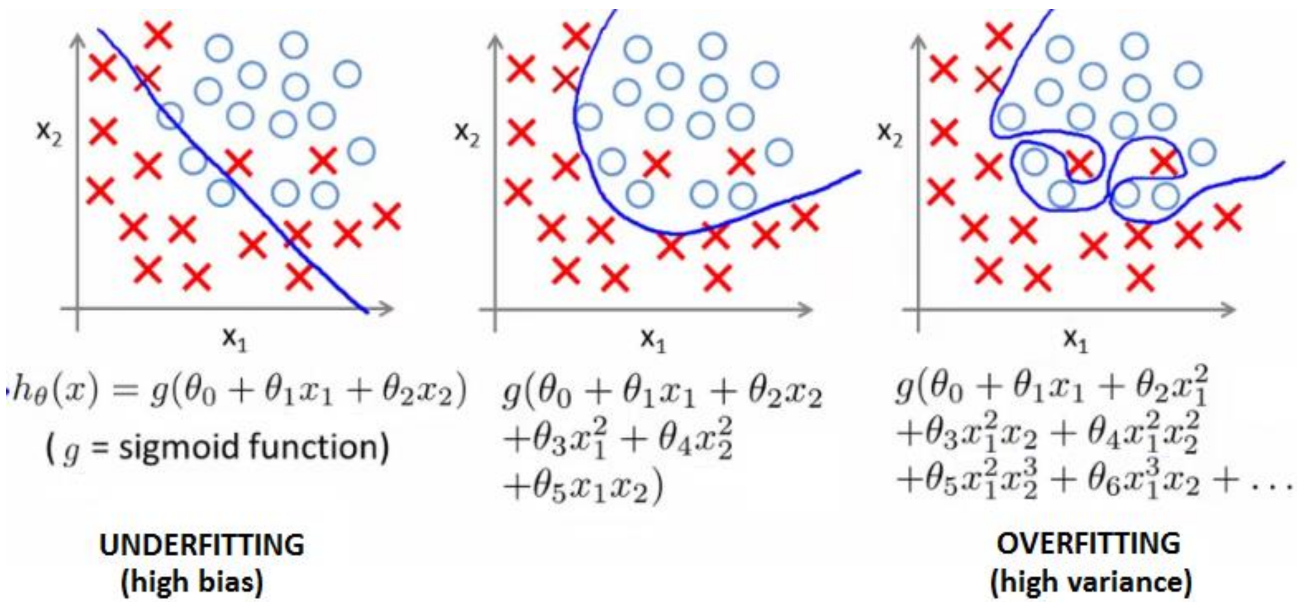
- Review of Linear Models
  - Linear Regression
  - Logistic Regression
  - Perceptron
- Generative vs. Discriminative
  - Hypothesis
  - Decision
  - Learning
- Over-fitting
  - ML - MAP
  - Regularization

Over-fitting

- Regression



- Classification



ML - MAP

- Maximum Likelihood (ML)

$$\boldsymbol{\theta}^*_{ML} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} p(X|\boldsymbol{\theta})$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{x} \epsilon X} \log p(\boldsymbol{x}|\boldsymbol{\theta})$$

likelihood

- Maximum A Posteriori (MAP)

$$\boldsymbol{\theta}^*_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|X) = \arg\max_{\boldsymbol{\theta}} \frac{p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(X)}$$

$$= \arg\max_{\boldsymbol{\theta}} p(X|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

likelihood • prior

$$= \arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{x} \in X} \log p(\boldsymbol{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

Regularization

- ML - MAP

$$\boldsymbol{\theta}^*_{ML} = \arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{x} \in X} \log p(\boldsymbol{x}|\boldsymbol{\theta})$$

$$\Downarrow$$

regularization term

$$\boldsymbol{\theta}^*_{MAP} = \arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{x} \in X} \log p(\boldsymbol{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

- Loss function plus regularization

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

$$\Downarrow$$

regularization term

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \hat{J}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) + \lambda R(\boldsymbol{\theta})$$

Example: Polynomial Curve Fitting

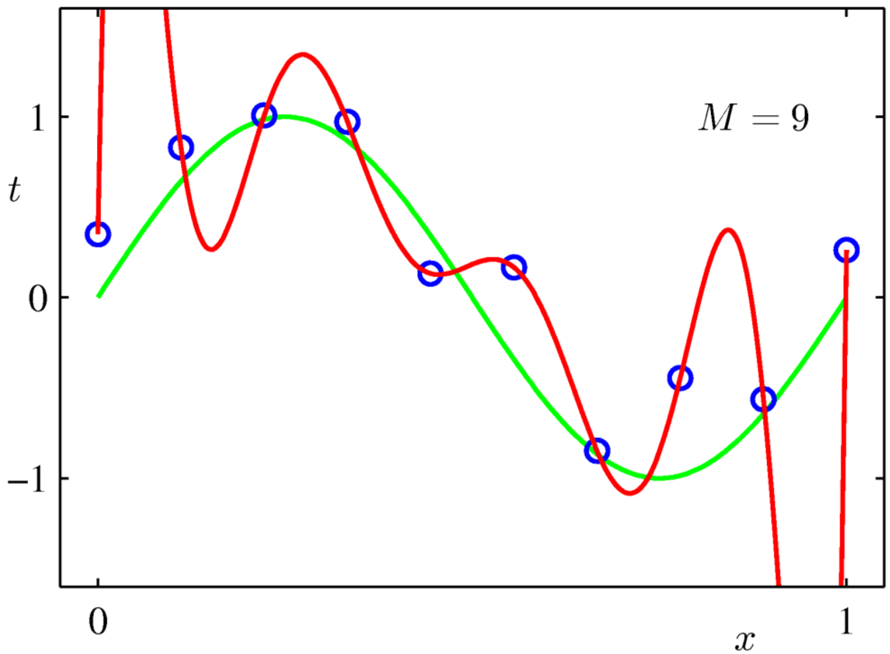
- ML (PRML Equation 1.62)

$$\ln \mathrm{p}(t|\boldsymbol{x},\boldsymbol{w},\beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(\boldsymbol{x}_n,\boldsymbol{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

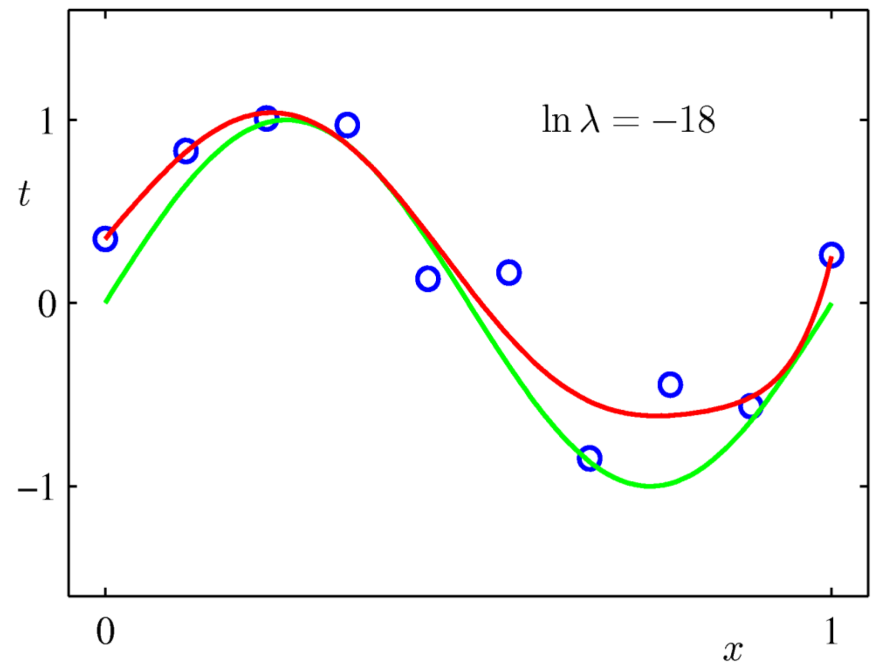
- MAP (PRML Equation 1.67)

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(\boldsymbol{x}_n,\boldsymbol{w}) - t_n\}^2 + \frac{\alpha}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$$

Example: Polynomial Curve Fitting



ML



MAP

Example: Logistic Regression

- ML

$$J_c(\boldsymbol{\theta}) = \sum_{i=1}^{N} y^{(i)} \log h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)\right)$$

$$where \;\; h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{1 + \exp^{-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}}}$$

- MAP

$$\hat{J}_c(\boldsymbol{\theta}) = \sum_{i=1}^{N} y^{(i)} \log h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)\right) + \frac{1}{2}||\boldsymbol{\theta}||^2$$

Example: Bernoulli Experiments

- Bernoulli Distribution

$$\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}, \qquad x \epsilon \{0,1\}$$

- Log-likelihood

$$\begin{aligned} \log L(\mu|X) &= \log \prod_{i=1}^{N} p(\boldsymbol{x}_i|\mu) = \sum_{i=1}^{N} \log p(\boldsymbol{x}_i|\mu) \\ &= m_1 \log p(1|\mu) + m_0 \log p(0|\mu) \\ &= m_1 \log \mu + m_0 \log(1-\mu) \end{aligned}$$

- ML solution

$$\frac{\partial \log L}{\partial \mu} = \frac{m_1}{\mu} - \frac{m_0}{1-\mu} = 0 \;\Leftrightarrow\; \hat{\mu}_{\mathrm{ML}} = \frac{m-1}{N}$$

Example: Bernoulli Experiments

- Prior Distribution

$$p(\mu|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1} \triangleq \mathrm{Beta}(\mu|\alpha,\beta)$$

$$where \quad \mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \qquad \Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$$

- MAP solution

$$\frac{\partial}{\partial\mu}\log L(\mu|X) + \log p(\mu) = \frac{m_1}{\mu} - \frac{m_0}{1-\mu} + \frac{\alpha-1}{\mu} - \frac{\beta-1}{1-\mu} = 0$$

$$\Leftrightarrow \hat{\mu}_{\mathrm{MAP}} = \frac{m_1+\alpha-1}{N+\alpha+\beta-2}$$

Outline

- Review of Linear Models
  - Linear Regression
  - Logistic Regression
  - Perceptron
- Generative vs. Discriminative
  - Hypothesis
  - Decision
  - Learning
- Over-fitting
  - ML - MAP
  - Regularization

Any Questions?