Bayesian Machine Learning: Some Basics

Feature Engineering Group

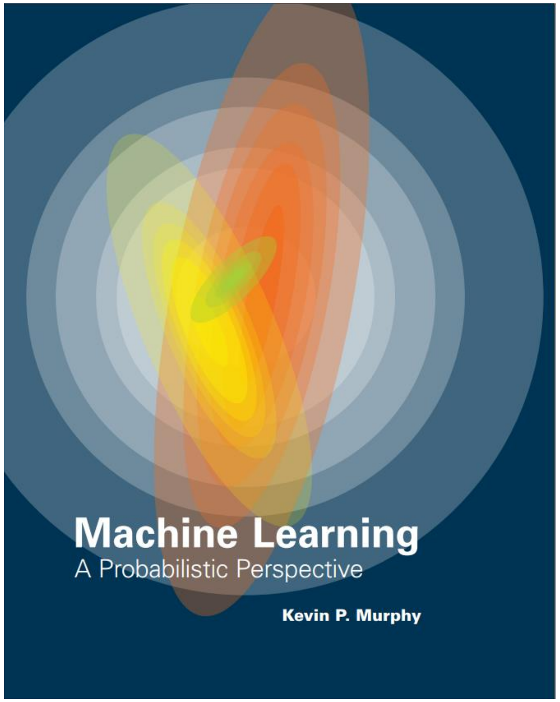
Chen Huang

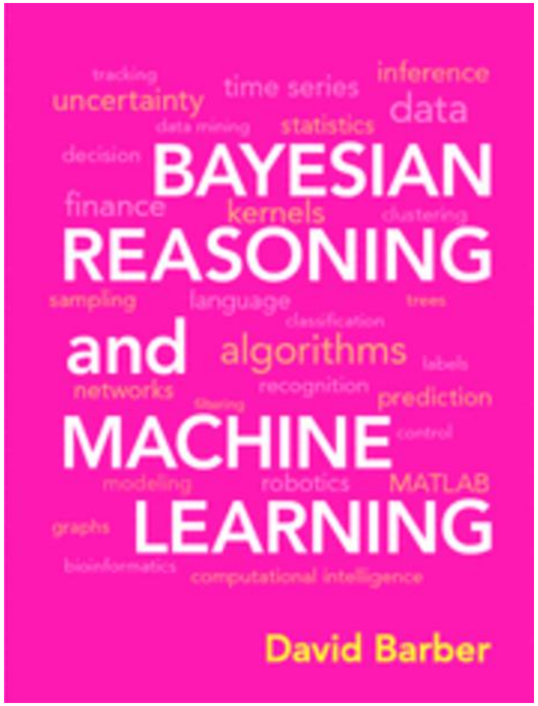
Feature Engineering Group, Data Mining Lab,
Big Data Research Center, UESTC
huangc.uestc@gmail.com

Reference

For more information

- http://fastml.com/bayesian-machine-learning/
- https://metacademy.org/roadmaps/rgrosse/bayesian_machine_learning





Reference

For more information

Research = search again and again…

Bayesian Machine Learning Contents

Gear Up

- ✓ Likelihood
- ✓ Prior
- ✓ Posterior
- ✓ Inference

Linear Regression

- ✓ Bayesian LR
- ✓ Prior & Regularizer
- ✓ Bayesian decision theory

Logistic Regression

- ✓ Bayesian LR
- ✓ Approximate inference
- ✓ Bayesian Model Selection

Gear Up

Four main steps

- **Likelihood** $P(D|\theta)$
  - Mechanism giving rise our observations D given a particular value of the parameters of interest
- **Prior** $P(\theta)$
  - Summarize our prior beliefs about the parameters
- **Posterior** $P(\theta|D)$
  - Using Bayes Theorem to combine prior beliefs with observed evidence
- **Inference** ***(Challenging problem)***
  - Use $P(\theta|D)$ to draw <u>further conclusions</u>
  - Algorithms: MAP/MCMC/VI

Gear Up!
Why Bayesian?

“ Further conclusions ”

- **Point estimation** if we must report a single best guess of theta
- **Make predictions** by averaging over the posterior distribution
- **Make decisions** so as to minimize posterior expected loss
- compare alternative models giving rise to **Bayesian model comparison**
- Naturally extend to **online** and **distributed learning**

Gear Up!
Bayesian or Not?

Some issues

- **How to choose a prior?**
  - Informative
  - Uninformative

- **Intractable integrals/posteriors**

$$P(D) = \int P(\theta)P(D|\theta)d\theta$$

  - Conjugate prior
  - Approximation →

- Gibbs sampling/MCMC
- Variational Inference
- Sequential MC/particle filter
- Stochastic MCMC/VI
- Streaming Variational Bayes

Gear Up!
Before we're going too far…

Give me some Bayesian models

- Gaussian Mixture Model
- Hidden Markov Model
- Conditional Random Field
- Bayesian Networks
- Gaussian Processes
- Dirichlet Process
- ……

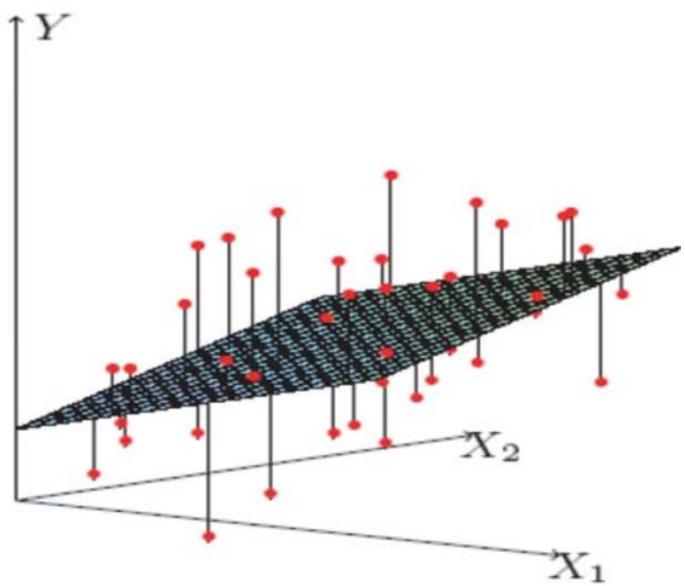
- BayesPA (*JMLR'14 @Jun Zhu*)
- OASIS (*AAAI'11 @Andrew B. Goldberg & Xiaojin Zhu*)

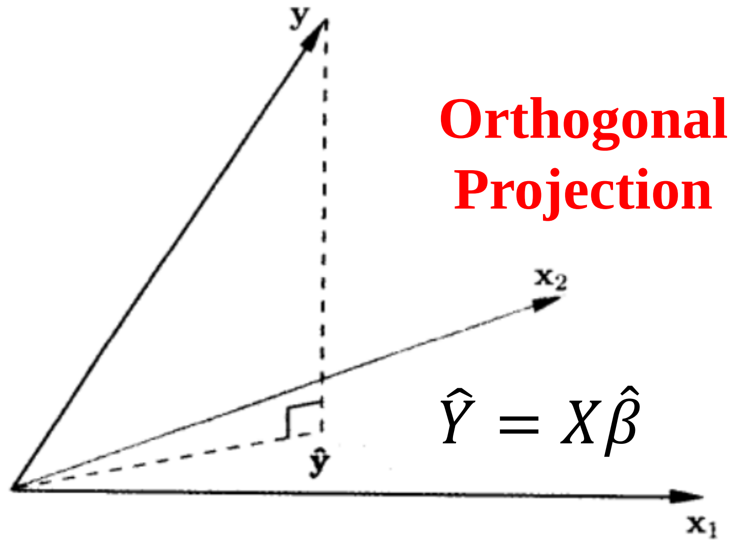
Linear Regression

Loss function

$$L\big(Y, F(X)\big) = L(Y, X\beta) = (Y - X\beta)^T (Y - X\beta)$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$





Orthogonal Projection

$$\hat{Y} = X\hat{\beta}$$

Linear Regression
From Bayesian Perspective

Four main steps

– **Likelihood** $P(D|\theta)$

$$y = f(X) + \varepsilon = XW + \varepsilon$$

– **Prior** $P(\theta)$

$$W \sim N(\mu, \Sigma);\ \varepsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

$$f \sim N(X\mu, X\Sigma X^T); \quad y \sim N(X\mu, X\Sigma X^T + \sigma^2 \boldsymbol{I})$$

– **Posterior** $P(\theta|D)$

$$P(W, y|X, \mu, \Sigma, \sigma^2) = N\left(\begin{bmatrix} \mu \\ X\mu \end{bmatrix}, \begin{bmatrix} \Sigma & (X\Sigma)^T \\ X\Sigma & X\Sigma X^T + \sigma^2 \boldsymbol{I} \end{bmatrix}\right)$$

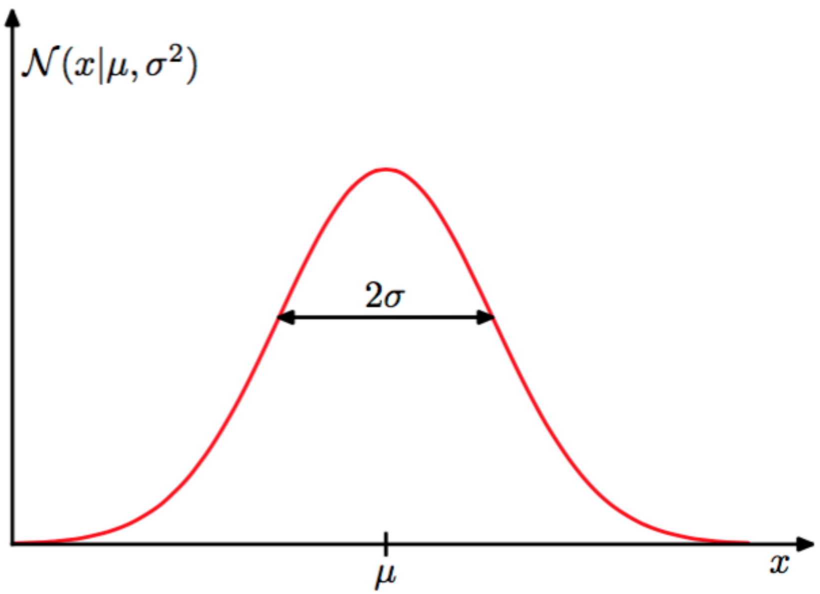
$$P(W|D, \mu, \Sigma, \sigma^2) = N(\mu_W, \Sigma_W)$$

Linear Regression
From Bayesian Perspective

MAP

– **Point estimation** for $P(\theta|D)$

$$P(W|D,\mu,\Sigma,\sigma^2) = N(\mu_W, \Sigma_W)$$



Linear Regression
From Bayesian Perspective

Bayesian predictive distribution

– **Predictive distribution** for $y^* = X^*W + \varepsilon$

$$P(y^*|D, X^*, \mu, \Sigma, \sigma^2)$$
$$= \int P(y^*|W, X^*, \sigma^2) P(W|D, \mu, \Sigma, \sigma^2) dW$$
$$= \mathrm{N}\left(\mathrm{X}^* \mu_\mathrm{W}, \mathrm{X}^* \Sigma_W {X^*}^T + \sigma^2 \boldsymbol{I}\right)$$

Make predictions by averaging over the posterior distribution of parameters

Why MAP?

*Why MAP

Bayesian decision theory

– Model **posterior expected loss** of a by averaging the loss function over the **unknown** parameter θ

$$\rho\big(p(\theta \mid \mathcal{D}), a\big) = \mathbb{E}\big[L(\theta, a) \mid \mathcal{D}\big] = \int_{\Theta} L(\theta, a) p(\theta \mid \mathcal{D})\, \mathrm{d}\theta.$$

How bad is my action

Weighted sum of loss caused by my action

*Why MAP

Bayesian decision theory → some facts

– The **Bayes estimator** of

A decision rule that minimizes posterior expected loss

– Posterior expected **squared loss** is posterior mean $\boldsymbol{E(\theta|D)}$
– Posterior expected **absolute loss** is posterior median
– Posterior expected (relaxed) **0-1 loss**, we have **MAP!**

Optimization is easier than integration!

*Why MAP

BDT for classification with 0-1 loss

– Bayes action is then to predict the class with the highest probability（MAP under 0-1 loss）

$$\mathbb{E}\left[L(y', a=1) \mid x', \mathcal{D}\right] = \Pr(y'=0 \mid x', \mathcal{D})$$
$$\mathbb{E}\left[L(y', a=0) \mid x', \mathcal{D}\right] = \Pr(y'=1 \mid x', \mathcal{D})$$

Bayesian Linear Regression

Relation to ridge regression

$w \sim N(\mu, s^2 I)$, $\mu = 0$, then it's the **ridge regression solution**

$$\boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}} = \boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^\top + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\mu})$$

$$\boldsymbol{\mu}_{\mathbf{w}|\mathcal{D}} = s^2\mathbf{X}^\top(s^2\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{y} = \left(\mathbf{X}^\top\mathbf{X} + \frac{\sigma^2}{s^2}\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

Regularization parameter

Mathematic Trick

$$(\mathbf{A}\mathbf{B} + c\mathbf{I})^{-1}\mathbf{A} = \mathbf{A}(\mathbf{B}\mathbf{A} + c\mathbf{I})^{-1}$$

Bayesian Linear Regression

Some facts

– **Ridge regression** = Bayesian linear regression with Gaussian prior on w and find the **MAP estimator**.

$$p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}, \sigma^2, s^2) \propto p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma^2) p(\mathbf{w} \mid s^2)$$

$$\begin{aligned}\hat{\mathbf{w}}_{\text{MAP}} &= \arg\max_{\mathbf{w}} -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\mathbf{x}_i^\top \mathbf{w} - y_i)^2 - \frac{1}{2s^2}\sum_{i=1}^{d} w_i^2 \\ &= \arg\min_{\mathbf{w}} \sum_{i=1}^{N}(\mathbf{x}_i^\top \mathbf{w} - y_i)^2 + \frac{\sigma^2}{s^2}\|\mathbf{w}\|_2^2,\end{aligned}$$

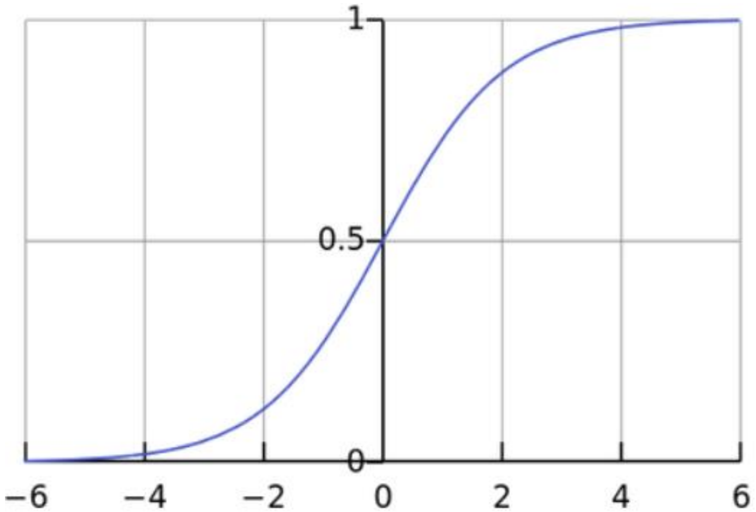
– **Sparsity = Laplacian priori on w**

Logistic Regression
From Linear Regression

Non-linear transformation

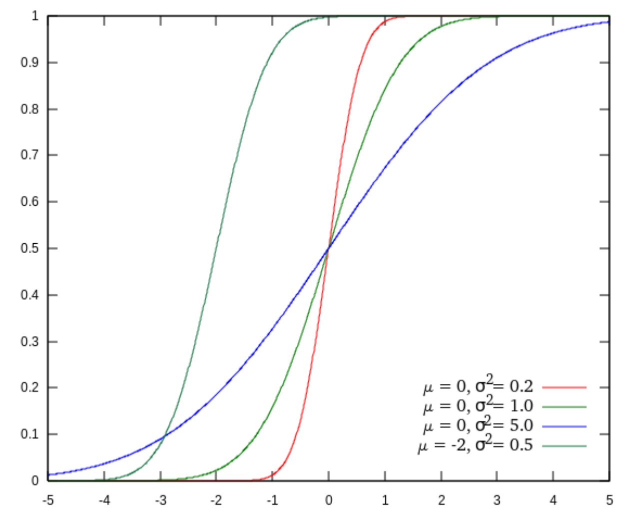
$$P(y = 1|X, W) = \boldsymbol{\sigma(XW)}$$

Logistic regression



$$\sigma(a) = \frac{\exp(a)}{1 + \exp(a)}$$

Probit regression



$$\sigma(a) = \Phi(a) = \int_{-\infty}^{a} \mathcal{N}(x; 0, 1^2)\,\mathrm{d}x.$$

Logistic Regression From Bayesian Perspective

Four main steps

- **Likelihood** $P(D|\theta)$ *(Bernoulli distribution)*

$$P(y|X,W) = \prod \sigma(X_i W)^{y_i}(1 - \sigma(X_i W))^{1-y_1}$$

- **Prior** $P(\theta)$

$$W \sim N(\mu, \Sigma);$$

- **Posterior** $P(\theta|D)$

$$P(W|D) = \frac{P(y|X,W)P(W)}{\int P(y|X,W)P(W)dW} = \frac{P(y|X,W)P(W)}{P(y|X)}$$

- **Inference**

Damn! Intractable

Intractable Posterior

How to solve

$$P(\theta|X) \approx Q(\theta)$$

- **Find an approximation to the posterior**
  - **Variational inference** $KL(P|Q)$
  - Laplacian Approximation $Q = N(\hat{\theta}, H)$
  - Assumed Density Filtering $KL(Q|\mathrm{P})$

- **Draw samples from the posterior**
  - Reject sampling
  - Importance sampling
  - **MCMC (MH, Gibbs)**
  - Slice sampling

Laplacian Approximation

Basic idea

- Target posterior

$$P(\theta|D) = \frac{P(D|\theta)\mathrm{P}(\theta)}{P(D)} = \frac{1}{Z}P(D|\theta)P(\theta)$$

- **Approximate $\boldsymbol{\delta(\theta)}$ by second-order taylor expansion.** **(*recall Newton methods*)**

$$\delta(\theta) = logP(D|\theta) + \log P(\theta)$$

$$\delta(\theta) \approx \delta(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^T H(\theta - \hat{\theta}) \qquad H = -\nabla^2\delta(\theta)\Big|_{\theta=\hat{\theta}}$$

$$P(\theta|D) \approx \exp\left(\delta(\hat{\theta})\right)\exp\left(-\frac{1}{2}(\theta - \hat{\theta})^T H(\theta - \hat{\theta})\right) = N(\hat{\theta}, H^{-1})$$

Laplacian Approximation
For Bayesian Logistic Regression

Basic idea

- **Predictive distribution**

$$P(y^*|X^*,D)$$

$$= \int \underbrace{P(y^*|W,X^*)}_{\sigma(XW)} \underbrace{P(W|D)}_{\approx\delta(W)} dW = \int \sigma(XW) N(\widehat{W}, H^{-1}) dW$$

- **Switch** $\sigma(XW)$

case *LOGISTIC_FUNCTION*

Still intractable

case *PROBIT_FUNCTION* **Bayesian Moderation**

$$\int \Phi(XW) N(\widehat{W}, H^{-1}) dW = \Phi\left(\frac{X^*\widehat{\mathrm{W}}}{\sqrt{1 + X^* H^{-1} {X^*}^T}}\right)$$

Laplacian Approximation and Bayesian Information Criterion

Basics about BIC

- A criterion for **Bayesian Model Selection**
- Model with the largest BIC is preferred
- Closely related to the *Akaike information criterion* (AIC)

- Given a set of models $\{M_i\}$, and observed data D

MAP estimator for M_i

Dimension of $\hat{\theta}_i$

$$BIC_i = logP\left(D\middle|\hat{\theta}_i\right) - \frac{d}{2}lnN$$

From Laplacian Approximation To BIC

Bayesian Model Selection

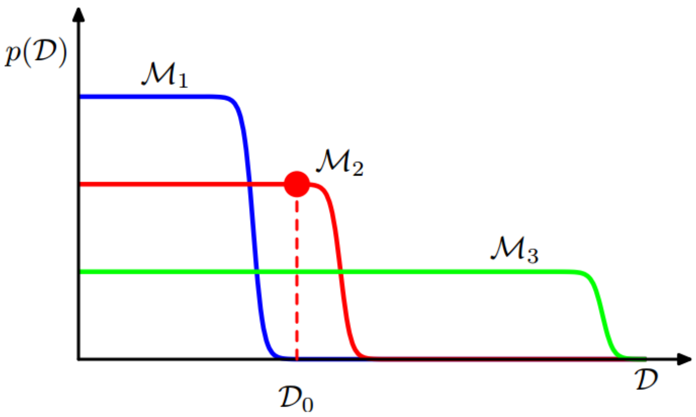
- **Model posterior**

$$P(M|D) = \frac{P(\mathrm{D}|\mathrm{M})P(M)}{P(D)}$$

- **Posterior odd**

Called ***"bayes factor*** in favor of M_i"

$$\frac{P(M_i|D)}{P(\mathrm{M_j}|\mathrm{D})} = \frac{P(\mathrm{D}|\mathrm{M}_i)P(M_i)}{P(\mathrm{D}|\mathrm{M}_j)P(M_j)} = \frac{P(M_i)\int P(D|\theta_i, M_i)P(\theta_i|M_i)d\theta_i}{P(M_j)\int P(D|\theta_j, M_j)P(\theta_j|M_j)d\theta_j}$$



Automatically gives a preference towards simpler models, in line with Occam's razor

From Laplacian Approximation To BIC

BIC

- **Posterior approximation**

$$P(\theta_i|D,M_i) = \frac{P(D|\theta_i,M_i)\mathrm{P}(\theta_i|M_i)}{\boxed{P(D|M_i)}} \approx N\left(\hat{\theta}_i, H^{-1}\right)$$

- **Model odd** *(under uniform distribution)*

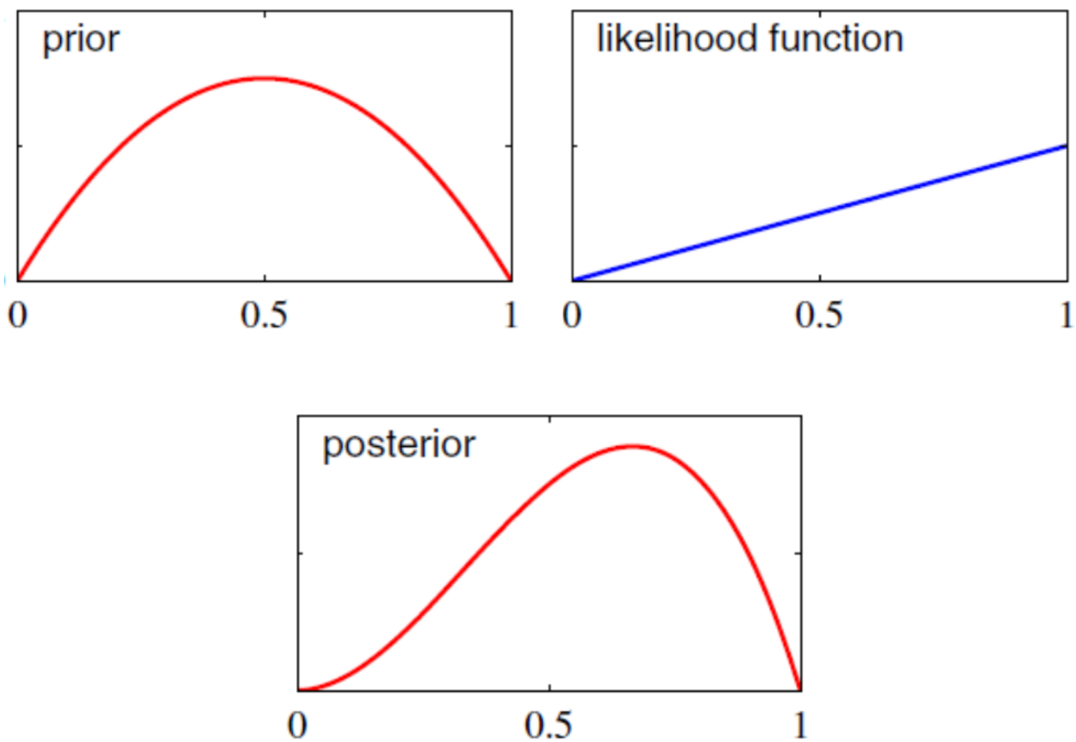
$$\frac{P(M_i|D)}{P(\mathrm{M_j}|\mathrm{D})} = \frac{\boxed{P(\mathrm{D}|\mathrm{M}_i)}}{P\left(\mathrm{D}\middle|\mathrm{M}_j\right)}$$

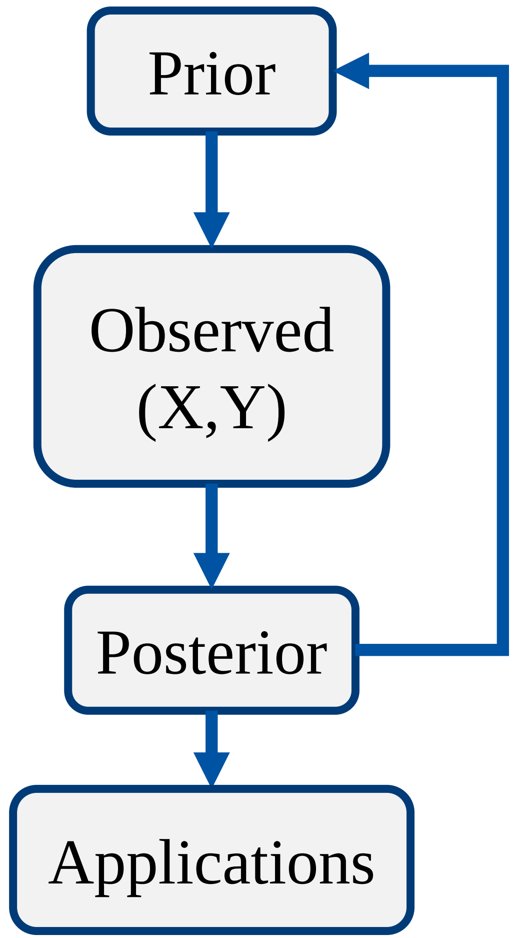
- BIC calculates $\log P(\mathrm{D}|\mathrm{M}_i)$ by Laplacian Approximation

$$P(\mathrm{D}|\mathrm{M}_i) \approx \int N\left(\hat{\theta}_i, H^{-1}\right) d\theta = \exp\left(\delta\left(\hat{\theta}_i\right)\right)\sqrt{\frac{(2\pi)^d}{|H|}}$$

Bayesian Online Learning

Sequential update



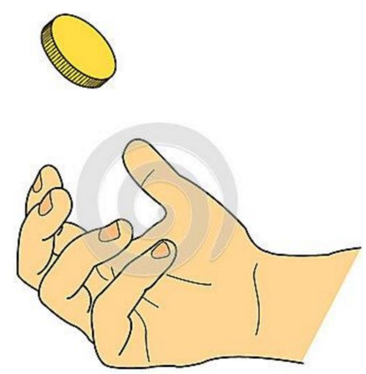


Bayesian Online Learning

Posterior Inference

– **Bayesian Conjugate**

> **Example**: Toss a coin
>
> Priori: $Beta(\alpha, \beta)$
>
> Likelihood: $Bernoulli(p)$
>
> Posteriori: $Beta(\alpha + heads, \beta + tails)$



– **Otherwise**

- MCMC / VI ?
- Stochastic variational inference (SVI)
- Sequential Monte Carlo (Particle filter)
- …..

Online Bayesian Passive-Aggressive Learning *(JMLR'14)*

What's the title means?

Online Learning
Based on Bayesian Framework
with Max Margin Property

Online
Bayesian
Passive-Aggressive Learning

Online Bayesian Passive-Aggressive Learning

Motivation

- **PA with Bayesian extension**
  - PA select one hyperplane (point estimation), which may be insufficient for some tasks (*latent variables.*)
- **Online version of Max-margin Bayesian**
  - Existing Max-margin Bayesian methods are offline

Online Learning
Recall PA & CW

Optimization function

– **PA**

$$W_{t+1} = \arg\min_{W} \frac{1}{2} \|W - W_t\|_2^2$$
$$s.t.\ L\big(W, (X_t, Y_t)\big) = 0$$

$$y = w \cdot x \quad w \sim N(\mu, \Sigma)$$

– **CW**

$$(u_{t+1}, \Sigma_{t+1}) = \min KL(N(u, \Sigma) || N(u_t, \Sigma_t))$$
$$s.t. P(Y_t(W \cdot X_i) \geq 0) \geq \eta$$

Online Bayesian Passive-Aggressive Learning

More general version of CW

$W \sim N(\mu, \Sigma)$	$(u_{t+1}, \Sigma_{t+1}) = \min KL(N(u, \Sigma) \| N(u_t, \Sigma_t))$ $s.t. P(Y_t(W \cdot X_i) \geq 0) \geq \eta$
$W \sim q(W)$	$\min \mathrm{KL}(q(W) \| q_t(W)) - E_{q(W)}[\log P(X_t \vert W)]$ $s.t. L_\varepsilon(q(W); (X_t, Y_t)) = 0$

More general assumption on the distribution of W

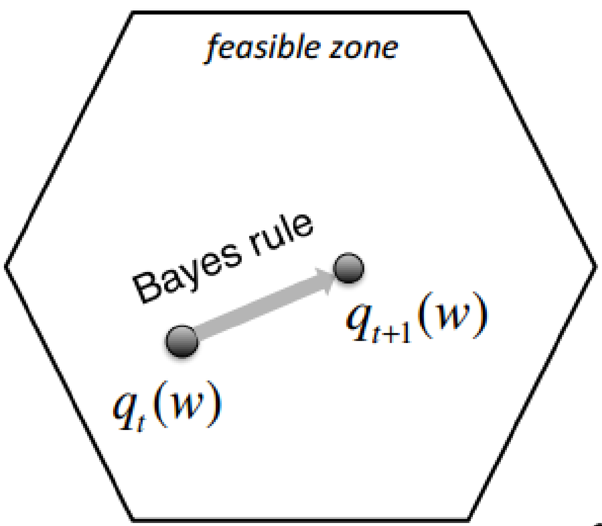
Seek large likelihood to serve new instance X_t

Online Bayesian Passive-Aggressive Learning

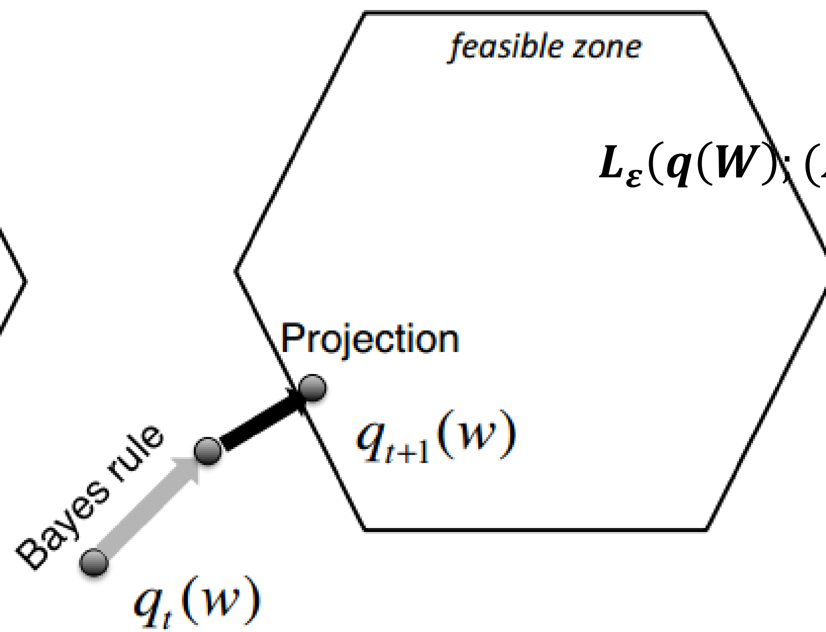
Passive-aggressive property

$$\min \int \mathrm{q(W)log}\frac{q(W)}{q_t(W)}dW - \int \log P(X_t|W)\, q(W)dW$$

$$= min \int \mathrm{q(W)log}\frac{q(W)}{q_t(W)P(X_t|W)}dW$$



$\boldsymbol{L_\varepsilon(q(W);(X_t,Y_t)) = 0}$



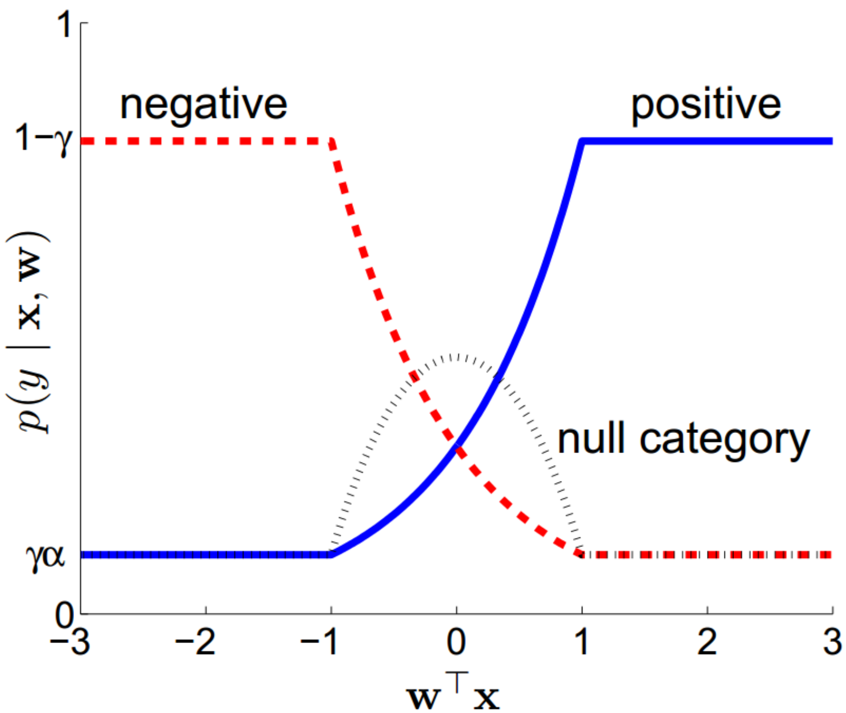
$\boldsymbol{L_\varepsilon(q(W);(X_t,Y_t)) \neq 0}$

Online Active Semi-Supervised Learning

(Bayesian framework, AAAI'11)

Motivation & Method

- General online Bayesian framework, which implements the **cluster assumption** through a special **likelihood**.
- Solved by sequential Monte Carlo with some algorithm to minimize particle degeneracy
- Buffer strategy to handle concept drift and achieve to be effective



Another Story…

New words

- Stochastic variational inference
- Sequential Monte Carlo / particle filters
- Probit regression
- Adversarial classification
- Maximum entropy discrimination
- Bayesian Monte Carlo
- ……

To Sum Up

Bayesian or not

- Bayesian ideas have had a big impact in machine learning in the past 20 years or so because of the **flexibility they provide in building structured models of real world phenomena.**

- A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule.

Gear Up
- ✓ Likelihood
- ✓ Prior
- ✓ Posterior
- ✓ Inference

Linear Regression
- ✓ Bayesian LR
- ✓ Prior & Regularizer
- ✓ Bayesian decision theory

Logistic Regression
- ✓ Bayesian LR
- ✓ Approximate inference
- ✓ Bayesian Model Selection

Thanks

By HC