Linear Regression Part III

CS 534: Machine Learning

Slides adapted from Lee Cooper, Joydeep Ghosh, Carlos Carvalho, and Ryan Tibshirani

Homework #1

- Posted and due Sep 19th at 11:59 PM on Gradescope
  - Single **typeset** PDF due on Gradescope HW1-Written with all the problems appropriately tagged
  - All code due on Gradescope HW1-Code (must include q2.py, elastic.py, and README.txt)

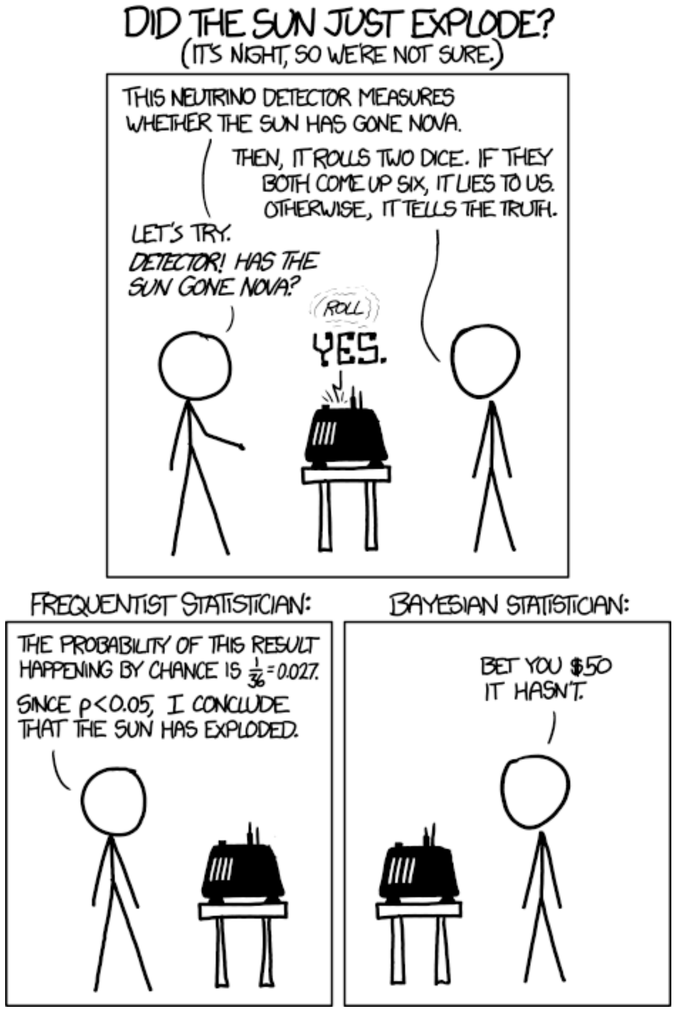
Review: Model Regularization

- Basic idea: Add penalty term on model parameters to achieve a more simple model or reduce sensitivity to training data

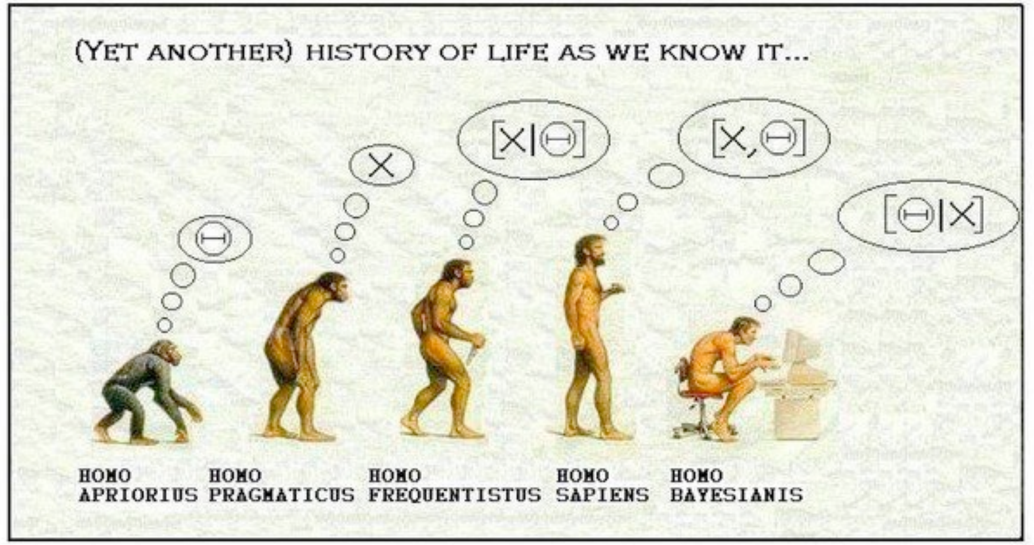
$$\min_{\beta} L(\mathbf{X}\boldsymbol{\beta}, \mathbf{y}) + \lambda \text{penalty}(\boldsymbol{\beta})$$

- Reasons:
  - Less prone to overfitting
  - Get the “right” model complexity

The War in Comics



http://www.xkcd.com/1132/



http://conversionxl.com/bayesian-frequentist-ab-testing/

Frequentist (“Classical") Statistics

- Data are repeatable random samples $D \sim p(y; \theta)$
- Probability model governs world we are observing
  - Parameters are fixed & constant
- Point estimation to estimate optimal parameter
- Prediction via the estimated parameter value

Binomial Experiment: Revisited

- Given a sequence of coin tosses x_1, x_2, …, x_M, we want to estimate the (unknown) probability of heads

$$P(H) = \theta$$

- Instances are i.i.d. samples

- Frequentist approach using MLE

$$\hat{\theta} = \frac{N_H}{N_H + N_T}$$

Is MLE the only option?

- Suppose that after 10 observations, MLE estimates the probability of a heads is 0.7, would you bet on heads for the next toss?
- How certain are you that the true parameter value is 0.7?
- Were there enough samples for you to be certain?

Bayesian Statistics

- Data are fixed and observed from the realized sample
- Parameters are random variables
- New “ingredient” is the prior distribution
  - Distribution reflecting belief about parameter
  - Chosen before being the data
- Prediction is expectation over unknown parameters

Bayesian Method

- Define the model
  - Prior distribution $p(\theta)$
  - Probability / likelihood model $p(D|\theta)$
- Compute posterior distribution using Bayes rule

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Bayesian Method: Interpretation

- Formulate knowledge about situation probabilistically
  - Model expresses qualitative aspects of our knowledge (e.g., forms of distributions, independence assumptions)
  - Prior represents beliefs about which values are more or less likely
- Posterior represents “rationally updated” beliefs after seeing data

Bayesian Method: Benefits

- No issue of justifying the estimator
- Only choices are the prior and likelihood function
- Posterior power
  - Reaching conclusions while accounting for uncertainty
  - Make predictions by averaging over posterior distribution

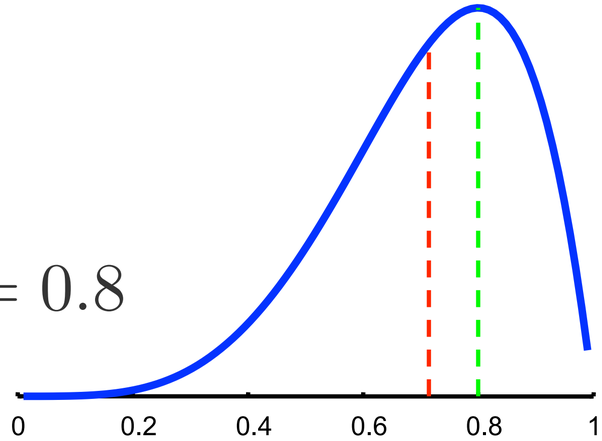
Uniform Prior

- Prior distribution: uniform for θ in [0, 1]
- Posterior distribution:

$$P(\theta|x_1, x_2, \cdots, x_M) \propto P(x_1, x_2, \cdots, x_M|\theta) \times 1$$

- Example: 5 coin tosses with 4 heads, 1 tail
  - MLE prediction:

$$P(\theta) = \frac{4}{5} = 0.8, P(x_{M+1} = H|D) = 0.8$$



  - Bayesian prediction:

$$P(x_{M+1} = H|D) = \int \theta P(\theta|D) d\theta = \frac{5}{7}$$

Beta Prior

- Prior distribution: $\theta \sim \text{Beta}(2, 2)$
- Posterior distribution:

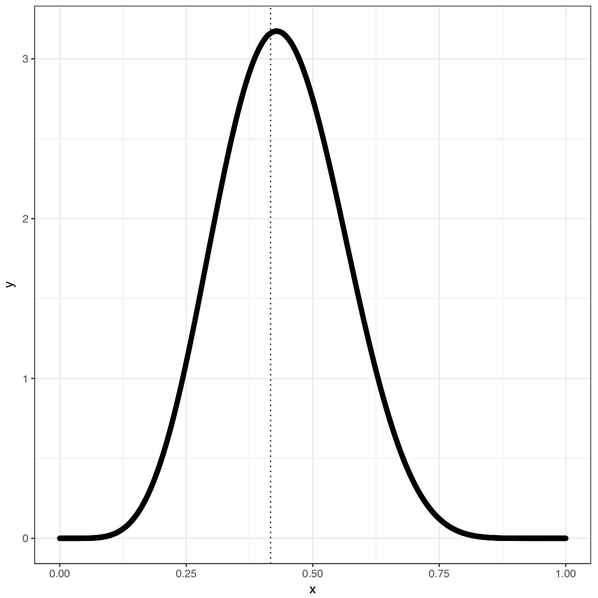
$$p(\theta|D) \propto \theta^{\alpha-1+n_h}(1-\theta)^{\beta-1+n_t} \sim \text{Beta}(n_h+\alpha, n_t+\beta)$$

- Example: 5 heads, 7 tails
  - MLE estimate:

$$\hat{\theta} = \frac{5}{5+7}$$

  - Bayesian estimate:

$$\theta|D \sim \text{Beta}(7, 9)$$



Posterior Distribution: Computation

- Analytical integration: works when “conjugate” prior distributions can be combined with likelihood — not often
- Gaussian approximation: posterior distribution close to Gaussian (Central Limit Theorem) — needs sufficient data compared to model complexity
- Markov Chain Monte Carlo: simulate a Markov chain that eventually converges to the posterior distribution (dominant)
- Variational approximation: cleverer way to approximate the posterior (sometimes faster) — not as general & exact

Bayesian Inference and MLE

- MLE and Bayesian prediction differ
- However…
  - IF prior is well-behaved (i.e., does not assign 0 density to any “feasible” parameter value)
  - THEN both MLE and Bayesian prediction converge to the same value as the number of training data increases

Limitations & Criticisms

- It is hard to come up with a prior (subjective) and the assumptions may be wrong
- Closed world assumption: need to consider all possible hypotheses for the data before observing the data
- Computationally demanding (compared to frequentist approach)
- Use of approximations weakens coherence argument

Bayesian Linear Regression

Bayesian Linear Regression

- What if we want to know the distribution over the coefficients and the variance?
- What if we want to quantify uncertainty?
- What if we have lots of outliers?
- What if we have a prior?

Review: Likelihood

- Assume observations with Gaussian noise

$$y = \boldsymbol{\beta}\mathbf{x} + \epsilon,\ \ p(\epsilon|\alpha) \sim N(\epsilon|0, \alpha^{-1})$$

$$\Updownarrow$$

$$p(y|\mathbf{x}, \boldsymbol{\beta}, \alpha) \sim N(y|\boldsymbol{\beta}\mathbf{x}, \alpha^{-1})$$

- Given observe data, likelihood function:

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \alpha) = \prod_{n=1}^{N} N(y_n|\boldsymbol{\beta}\mathbf{x}_n, \alpha^{-1})$$

Bayesian Approach: Simple Case

- Define a conjugate prior over coefficients

$$p(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}|\mathbf{m}_0, \mathbf{S}_0)$$

- Combine with likelihood function

$$p(\boldsymbol{\beta}|\mathbf{y}) \sim N(\boldsymbol{\beta}|\mathbf{m}_N, \mathbf{S}_N)$$
$$\mathbf{m}_N = S_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \alpha \mathbf{X}^\top \mathbf{y}\right)$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \alpha \mathbf{X}^\top \mathbf{X}$$

Bayesian Approach: Common Case

- Common choice for prior

$$p(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}|0, \kappa^{-1}\mathbf{I})$$

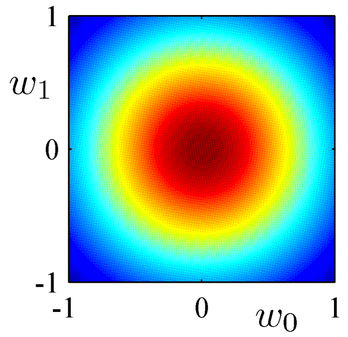
- Posterior distribution

$$\mathbf{m}_N = \alpha \mathbf{S}_N \mathbf{X}^\top \mathbf{y}$$
$$\mathbf{S}_N^{-1} = \kappa \mathbf{I} + \alpha \mathbf{X}^\top \mathbf{X}$$

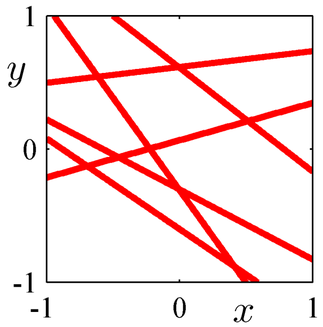
Example: Bayesian LR

0 data points observed

Prior



Data Space



Example: Bayesian LR