Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 5 - Linear Regression - Bayesian Inference and Regularization

Building on questions on Least Squares Linear Regression

1. Is there a probabilistic interpretation?
    - Gaussian Error, Maximum Likelihood Estimate
2. Addressing overfitting
    - Bayesian and Maximum Aposteriori Estimates, Regularization
3. How to minimize the resultant and more complex error functions?
    - Level Curves and Surfaces, Gradient Vector, Directional Derivative, Gradient Descent Algorithm, Convexity, Necessary and Sufficient Conditions for Optimality

Recap: Illustration through a Simple Coin Tossing Example: Maximum Likelihood Estimation vs. Bayesian Estimation

Recap: Case Study

Suresh likes to toss coins. One day he decided to count the number of heads and tails in his coin tosses. Here is what he found. After tossing 1000 times (it took him a hours, but he likes to toss coins), he found that the coin landed on heads 400 times and tails 600 times. His reflection: If I were to toss the coin once more time, what is the probability that I get a heads?

Recap: MLE estimate for Coin Tossing

- We restate Suresh's problem as the MLE of the probability of getting a head. This is the value of p which maximizes the likelihood of observing 400 heads as outcomes.

$$\hat{p} = \underset{p}{\operatorname{argmax}} \ {}^{1000}C_{400}p^{400}(1-p)^{600}$$

- $\hat{p} = 0.4$ as we had intuitively guessed. In general, the value of p which maximises the likelihood of observing h heads, given n coin tosses is.

$$\hat{p} = \underset{p}{\operatorname{argmax}} \ {}^{n}C_{h}p^{h}(1-p)^{n-h}$$

Bayesian Inference/Estimation

Case Study:

Suresh now brings a newly minted coin to toss. He *believes* that the coin is fair and heads and tails are equally likely outcomes (since the coin is not worn out). Now like always he flips the coin 4 times, and finds out that heads appeared all the 4 times.

1. Is the MLE estimate $\hat{p} = 1$ intuitive? Is tails *improbable*?
2. Is there a way that Suresh could update his *belief* about the coin.

Bayesian Inference

- H: One of few competing hypotheses whose probability may be affected by observed data.
- $\Pr(H)$: The (prior) probability of H before data $\mathcal{D}$ is observed. This indicates one's previous *belief* in the hypothesis.
- The evidence $\mathcal{D}$: New data that were not used in computing the prior probability

$$p(H \mid \mathcal{D}) \propto p(\mathcal{D} \mid H)\ p(H)$$

Conjugate Prior

Let $\mathcal{D} \mid H$ follow a distribution d_1 and H follow a distribution d_2. The distribution d_2 is the conjugate prior of d_1 if the distribution of $\Pr(H \mid \mathcal{D})$ follows the distribution d_2.

Some Examples:

1. (d_1) Bernoulli & Binomial - Beta (d_2)
2. Geometric - Beta
3. Categorical - Dirichlet
4. Multinomial - Dirichlet
5. Poisson - Gamma
6. Normal - Inverse Gamma

The Beta Conjugate Prior for Bernoulli/Binomial

Recall form of binomial: ${}^{n}C_{h}\, p^{h}(1-p)^{n-h}$ $(n-h=t)$

${}^{h+t}C_{h}\, p^{h}(1-p)^{t}$

Let $\mathcal{D} \mid H$ follow a distribution $Ber(p)$ (p is probability of heads) and p follow a distribution $Beta(p;\alpha,\beta) \sim \frac{p^{(\alpha-1)}(1-p)^{(\beta-1)}}{B(\alpha,\beta)}$,

- *The beta normalization function:*

$$P(H) = B(\alpha,\beta) = \int_{p=0}^{1} p^{(\alpha-1)}(1-p)^{(\beta-1)}dp = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$ where $\Gamma(.)$ behaves like the factorial function: $\Gamma(n) = (n-1)!$ if $n \in \mathbb{Z}^+$

If $\alpha, \beta \in \mathbb{Z}^+$ $B(\alpha,\beta) = \frac{1}{{}^{\alpha+\beta-1}C_{\alpha-1} \times \beta}$

i.e. $\frac{1}{B(\alpha,\beta)} = {}^{\alpha+\beta-1}C_{\alpha-1} \times \beta$ → Similar to ${}^{h+t}C_{h}$

The Beta Conjugate Prior for Bernoulli/Binomial

p is a param

Let $\mathcal{D} \mid H$ follow a distribution $Ber(p)$ (p is probability of heads)
and p follow a distribution $Beta(p;\alpha,\beta) \sim \frac{p^{(\alpha-1)}(1-p)^{(\beta-1)}}{B(\alpha,\beta)}$,

- $\Pr(H \mid \mathcal{D}) = \Pr(p \mid \mathcal{D}) = \frac{p(x_1 \ldots x_n \mid p)\, p(p)}{\int_p p(x_1 \ldots x_n \mid p)\, p(p)}$

we know from conjugacy that this MUST be a Beta distribution

$$= \frac{{}^nC_h\, p^h (1-p)^{n-h} * p^{(\alpha-1)} (1-p)^{\beta-1}}{\int_p \cdots}$$

$$= p^{h+\alpha-1} (1-p)^{n-h+\beta-1} + \cdots$$

$$= Beta(p; \alpha+h, \beta+n-h)$$

The Beta Conjugate Prior for Bernoulli/Binomial

Let $\mathcal{D} \mid H$ follow a distribution $Ber(p)$ (p is probability of heads) and p follow a distribution $Beta(p; \alpha, \beta) \sim \frac{p^{(\alpha-1)}(1-p)^{(\beta-1)}}{B(\alpha,\beta)}$,

- $\Pr(H \mid \mathcal{D}) = \Pr(p \mid \mathcal{D}) = \dfrac{\Pr(\mathcal{D}|p)\Pr(p)}{\displaystyle\int_q \Pr(\mathcal{D} \mid q)\Pr(q)}$

$$= \frac{{}^nC_h p^h (1-p)^{n-h} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{(\alpha-1)}(1-p)^{(\beta-1)}}{\displaystyle\int_q {}^nC_h q^h (1-q)^{n-h} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} q^{(\alpha-1)}(1-q)^{(\beta-1)}}$$

$$\propto p^{\alpha+h-1}(1-p)^{\beta+n-h-1} \sim \underline{Beta(p;}\, \alpha + \underline{h}, \beta + \underline{n-h})$$

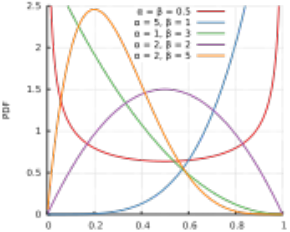
Added # heads to α &
# tails to β

More on the $Beta(\alpha, \beta)$ distribution

1. Mean $\mathbf{E}_{Beta(\alpha,\beta)}[p] = \boxed{\frac{\alpha}{\alpha+\beta}}$ and $\arg\max_p \; Beta(p;\alpha,\beta) = \frac{\alpha-1}{\alpha+\beta-2}$ Mode (the mode of the distribution)

$$\approx \frac{\#\text{prior heads}}{\#\text{prior tosses}}$$

2. $Beta(1,1)$ is the uniform distribution!
3. Is the conjugate prior pdf for the Bernoulli, binomial, negative binomial and geometric distributions and has the following pdf:



The MAP Estimate for Bernoulli/Binoimal

Let $\mathcal{D} \mid H$ follow a distribution $Ber(p)$ (p is probability of heads) and p follow a distribution $Beta(p;\alpha,\beta) \sim \frac{p^{(\alpha-1)}(1-p)^{(\beta-1)}}{B(\alpha,\beta)}$,

1. *The Maximum Likelihood Estimate:*

$$\underline{\hat{p}} = \underset{p}{\operatorname{argmax}}\ {}^{n}C_{h}p^{h}(1-p)^{n-h} = \boxed{\frac{h}{n}}$$

2. *The Maximum a-Posterior (MAP) Estimate:* The mode of the posterior distribution

$$\underline{\tilde{p}} = \underline{\underset{H}{\operatorname{argmax}}\ \Pr(H \mid \mathcal{D})} = \underline{\underset{p}{\operatorname{argmax}}\ \Pr(p \mid \mathcal{D})} = \frac{\alpha+h-1}{\alpha+\beta+n-2}$$

Recall: $\underset{p}{\operatorname{argmax}}\ Beta(p;\alpha,\beta) = \frac{\alpha-1}{\alpha+\beta-2}$

The MAP Estimate for Bernoulli/Binoimal

Let $\mathcal{D} \mid H$ follow a distribution $Ber(p)$ (p is probability of heads) and p follow a distribution $Beta(p;\alpha,\beta) \sim \frac{p^{(\alpha-1)}(1-p)^{(\beta-1)}}{B(\alpha,\beta)}$,

1. *The Maximum Likelihood Estimate:*

   $$\hat{p} = \underset{p}{\operatorname{argmax}}\ {}^{n}C_{h} p^{h}(1-p)^{n-h} = \frac{h}{n}$$

2. *The Maximum a-Posterior (MAP) Estimate:* The mode of the posterior distribution

   $$\tilde{p} = \underset{H}{\operatorname{argmax}}\ \Pr(H \mid \mathcal{D}) = \underset{p}{\operatorname{argmax}}\ \Pr(p \mid \mathcal{D})$$

   $$= \underset{p}{\operatorname{argmax}}\ Beta(p;\alpha+h,\beta+n-h) = \underline{\frac{\alpha+h-1}{\alpha+\beta+n-2}}$$

Case Study Continued

Coming back to the Suresh's case study, he observed 4 heads on 4 tosses, his MLE is

$$\underline{\hat{p}} = \underset{p}{\operatorname{argmax}}\ {}^4C_4 p^4 (1-p)^0 = \underline{1}$$

If his prior on p was $Beta(p; \overset{\alpha=3}{3}, \overset{\beta=3}{3})$, then his posterior will be $Beta(p; 3+4, 3+0) = Beta(p; 7, 3)$ and his MAP estimate will be

$$\underline{\hat{p}} = \underset{p}{\operatorname{argmax}}\ Beta\underline{(p; 7, 3)} = \frac{7-1}{7+3-2} = 0.75$$

Often, α, β are tuned to maximize performance on validation / held-out data

Summary: Bayesian Inference for Bernoulli

Let $\mathcal{D} \mid H$ follow a distribution $Ber(p)$ (p is probability of heads) and p follow a distribution $Beta(p;\alpha,\beta) \sim \frac{p^{(\alpha-1)}(1-p)^{(\beta-1)}}{B(\alpha,\beta)}$,

1. *The Maximum Likelihood Estimate:*
   $$\hat{p} = \operatorname*{argmax}_{p} \; {}^{n}C_{h} p^{h}(1-p)^{n-h} = \frac{h}{n}$$
2. *The Posterior Distribution:*
   $$\Pr(p \mid \mathcal{D}) = Beta(p;\alpha+h,\beta+n-h)$$
3. *The Maximum a-Posterior (MAP) Estimate:* The mode of the posterior distribution
   $$\tilde{p} = \operatorname*{argmax}_{H} \Pr(H \mid \mathcal{D}) = \operatorname*{argmax}_{p} \Pr(p \mid \mathcal{D})$$
   $$= \operatorname*{argmax} Beta(p;\alpha+h,\beta+n-h) = \frac{\alpha+h-1}{\alpha+\beta+n-2}$$

Illustration of Bayesian Estimation on a continuous valued random variable

Conjugate Prior for (univariate) Gaussian

- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 \ldots x_m$
- $\mu_{MLE} = \frac{1}{m}\sum_{i=1}^{m} x_i$ and $\sigma^2_{MLE} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{MLE})^2$ $\Big\}\ \hat{\mu}, \hat{\sigma} = \underset{\mu,\sigma}{\arg\max} \prod_i N(x_i|\mu,\sigma)$
- Suppose you are told that the conjugate prior for the (univariate) normally distributed random variable X in the case that σ^2 is not a random variable is (Prior belief on μ, NOT on σ^2) $\underline{\Pr(\mu)} = \underline{\mathcal{N}(\mu_0, \sigma_0^2)}$. Then the **posterior** is?

$$\Pr(H|D) = \Pr(\mu|D) = \mathcal{N}(\mu_n, \sigma_n^2)$$

How will you find μ_n & σ_n^2? Trick??

$$P(x_1 .. x_n \mid \mu, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(\sum_{i=1}^{n} \frac{(x_1-\mu)^2}{2\sigma^2}\right)$$

$$\mu_{MLE}, \sigma^2_{MLE} = \underset{\mu,\sigma^2}{\arg\max} \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\mu_{MLE} = \underbrace{\underset{\mu}{\arg\max} -\sum_{i=1}^{n} (x_1-\mu)^2}_{}$$

(Intuitively, μ_{MLE} is independent of σ^2)

Recall, mean mean minimizes sum of squares of deviation $\Rightarrow \mu_{MLE} = \frac{1}{n}\sum_i x_i$

Conjugate Prior for (univariate) Gaussian

- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 \ldots x_m$
- $\mu_{MLE} = \frac{1}{m}\sum_{i=1}^{m} x_i$ and $\sigma^2_{MLE} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{MLE})^2$
- Suppose you are told that the conjugate prior for the (univariate) normally distributed random variable X in the case that σ^2 is not a random variable is $\Pr(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$. Then the **posterior** is?
- Answer: $\Pr(\mu|x_1 \ldots x_m) = \mathcal{N}(\mu_m, \sigma_m^2)$ such that $\mu_m = \ldots\ldots$ and $\dfrac{1}{\sigma_m^2} = \ldots..$
- Helpful tip: Product of Gaussians is always a Gaussian

$$\text{①}\ \Pr(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} exp\left(\frac{-(\mu-\mu_0)^2}{2\sigma_0^2}\right) \sim N(\mu_0, \sigma_0^2)$$

$$\Pr(x_i|\mu;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) \sim N(\mu, \sigma^2)$$

$$\text{②}\ \Pr(\mathcal{D}|\mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^m exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2\right)$$

$$\Pr(\mu|\mathcal{D}) \propto \Pr(\mathcal{D}|\mu)\Pr(\mu) = \text{①} * \text{②}$$

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^m \frac{1}{\sqrt{2\pi\sigma_0^2}} exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \propto$$

$$exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) = \underbrace{exp\left(\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2\right)}_{N(\mu_m, \sigma_m^2)}$$

Detailed derivation (contd.)

from Pr(D|M)

from Pr(M)

Our reference equality:

$$exp\left(\underbrace{\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2}_{} - \underbrace{\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}_{}\right) = exp\left(\underbrace{\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2}_{}\right),$$

Since we believe that Pr(M|D) follows Normal distr

Matching coefficients of μ^2, we get

$$\frac{-1}{2\sigma^2}\left(m\cancel{\mu^2}\right) - \frac{\cancel{\mu^2}}{2\sigma_0^2} = \frac{-1}{2\sigma_m^2}\cancel{\mu^2}$$

Detailed derivation (contd.)

Our reference equality:

$$exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)=exp\left(\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2\right),$$

Matching coefficients of μ^2, we get

$\frac{-\mu^2}{2\sigma_m^2}=\frac{-\mu^2}{2}(\frac{m}{\sigma^2}+\frac{1}{\sigma_0^2})\Rightarrow$

Detailed derivation (contd.)

Our reference equality:

$$exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)=exp\left(\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2\right),$$

Matching coefficients of μ^2, we get

$$\tfrac{-\mu^2}{2\sigma_m^2}=\tfrac{-\mu^2}{2}(\tfrac{m}{\sigma^2}+\tfrac{1}{\sigma_0^2})\Rightarrow\frac{1}{\sigma_m^2}=\frac{1}{\sigma_0^2}+\frac{m}{\sigma^2}$$

Matching coefficients of μ, we get

Detailed derivation (contd.)

Our reference equality:

$$exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)=exp\left(\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2\right),$$

Matching coefficients of μ^2, we get

$\frac{-\mu^2}{2\sigma_m^2}=\frac{-\mu^2}{2}(\frac{m}{\sigma^2}+\frac{1}{\sigma_0^2})\Rightarrow\dfrac{1}{\sigma_m^2}=\dfrac{1}{\sigma_0^2}+\dfrac{m}{\sigma^2}$

Matching coefficients of μ, we get

$\frac{2\mu\mu_m}{2\sigma_m^2}=\mu\left(\frac{2\sum_{i=1}^{m}x_i}{2\sigma^2}+\frac{2\mu_0}{2\sigma_0^2}\right)\Rightarrow$

Detailed derivation (contd.)

Our reference equality:

$$exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)=exp\left(\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2\right),$$

Matching coefficients of μ^2, we get

$$\frac{-\mu^2}{2\sigma_m^2}=\frac{-\mu^2}{2}\left(\frac{m}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\Rightarrow\frac{1}{\sigma_m^2}=\frac{1}{\sigma_0^2}+\frac{m}{\sigma^2}$$

Matching coefficients of μ, we get

$$\frac{2\mu\mu_m}{2\sigma_m^2}=\mu\left(\frac{2\sum_{i=1}^m x_i}{2\sigma^2}+\frac{2\mu_0}{2\sigma_0^2}\right)\Rightarrow\mu_m=\sigma_m^2\left(\frac{\sum_{i=1}^m x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\text{ or}$$

$$\mu_m=\sigma_m^2\left(\frac{m\hat{\mu}_{ML}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\Rightarrow$$

Detailed derivation (contd.)

Our reference equality:

$$exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)=exp\left(\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2\right),$$

Matching coefficients of μ^2, we get

$$\frac{-\mu^2}{2\sigma_m^2}=\frac{-\mu^2}{2}(\frac{m}{\sigma^2}+\frac{1}{\sigma_0^2})\Rightarrow\frac{1}{\sigma_m^2}=\frac{1}{\sigma_0^2}+\frac{m}{\sigma^2}$$

Matching coefficients of μ, we get

$$\frac{2\mu\mu_m}{2\sigma_m^2}=\mu\left(\frac{2\sum_{i=1}^m x_i}{2\sigma^2}+\frac{2\mu_0}{2\sigma_0^2}\right)\Rightarrow\mu_m=\sigma_m^2\left(\frac{\sum_{i=1}^m x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)$$ or

$$\mu_m=\sigma_m^2\left(\frac{m\hat{\mu}_{ML}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\Rightarrow\mu_m=(\frac{\sigma^2}{m\sigma_0^2+\sigma^2}\mu_0)+(\frac{m\sigma_0^2}{m\sigma_0^2+\sigma^2}\hat{\mu}_{ML})$$