Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 4 - Bayesian Estimation and Bayesian Linear Regression

Estimating $\mathbf{w}$: Maximum Likelihood

- If $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $y = \mathbf{w}^{\mathbf{T}}\phi(\mathbf{x}) + \epsilon$ where $\mathbf{w},\ \phi(\mathbf{x}) \in \mathbf{R^m}$ then, given dataset $\mathcal{D}$, find the most likely $\hat{\mathbf{w}_{ML}}$

Error that prevents a perfect linear fit

- Recall: $\Pr(y_j|\mathbf{x}_j, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{(y_j - \mathbf{w}^T\phi(\mathbf{x}_j))^2}{2\sigma^2}\right) = \mathcal{N}(\mathbf{w}^T\phi(\mathbf{x}), \sigma^2)$

Data = dependent variable(s)

- From *Probability of data* to *Likelihood of parameters*: $\Pr(\mathcal{D}|\mathbf{w}) = L(\mathbf{w}|\mathcal{D}) = \Pr(\mathbf{y}|\mathbf{x}, \mathbf{w}) =$

$\hat{w} = \underset{w}{\operatorname{argmax}}\ L(w|D)$

$$\Pr(y_1, y_2 \ldots y_m | w, x_1, x_2 \ldots) = \prod_j \Pr(y_j | w, x_j) = \prod_j \mathcal{N}(w^T\phi(x_j), \sigma^2)$$

each ϵ_j is independent & has same distr

since linear fns of independent r.vs are independent

since all y_j's are identically distributed

Often, $\epsilon_1, \epsilon_2 \ldots \epsilon_m$ or $y_1 \ldots y_m$ are referred to as iid samples

iid: independent identically distributed

Estimating $\mathbf{w}$: Maximum Likelihood

- If $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $y = \mathbf{w}^{\mathbf{T}}\phi(\mathbf{x}) + \epsilon$ where $\mathbf{w},\ \phi(\mathbf{x}) \in \mathbf{R}^{\mathbf{m}}$ then, given dataset $\mathcal{D}$, find the most likely $\hat{\mathbf{w}_{ML}}$
- Recall: $\Pr(y_j|\mathbf{x}_j, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{(y_j - \mathbf{w}^T\phi(\mathbf{x}_j))^2}{2\sigma^2}\right)$
- From *Probability of data* to *Likelihood of parameters*: $\Pr(\mathcal{D}|\mathbf{w}) = L(\mathbf{w}|\mathcal{D}) = \Pr(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \prod_{j=1}^{m} \Pr(y_j|\mathbf{x}_j, \mathbf{w}) = \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{(y_j - \mathbf{w}^T\phi(\mathbf{x}_j))^2}{2\sigma^2}\right)$
- Maximum Likelihood Estimate $\hat{\mathbf{w}}_{ML} = \underset{\mathbf{w}}{\text{argmax}}\ \Pr(\mathcal{D}|\mathbf{w}) = \underset{\mathbf{w}}{\text{argmax}}\ L(\mathbf{w}|\mathcal{D})$

Joint distrib of $\mathcal{D}$ given $\mathbf{w}$

Viewing $p_{\mathbf{w}}(\mathcal{D}|\mathbf{w})$ as a fn of $\mathbf{w}$ for given $y_1 \cdots y_m$ in $\mathcal{D}$

Optimization Trick

- Optimization Trick: Optimal point is invariant under monotonically increasing transformation (such as log) [Tutorial 2]

$$L(w|D) = \prod_j \frac{1}{\sqrt{\sigma^2 2\pi}} \exp\left(-\frac{(y_j - w^T\phi(x_j))^2}{2\sigma^2}\right)$$

$$\hat{w}_{MLE} = \arg\max_w L(w|D) = \arg\max_w \log\left(L(w|D)\right)$$

Optimization Trick

- Optimization Trick: Optimal point is invariant under monotonically increasing transformation (such as log) [Tutorial 2]
- $\log L(\mathbf{w}|\mathcal{D}) = LL(\mathbf{w}|\mathcal{D}) =$

$$-\frac{m}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{m}(\mathbf{w}^T\phi(\mathbf{x_j}) - \mathbf{y_j})^2$$

For a fixed (yet unknown) σ^2

$$\hat{\mathbf{w}}_{ML} = \text{argmax}_{\mathbf{w}} \frac{-1}{2\sigma^2}\sum_{j=1}^{m}(\mathbf{w}^T\phi(\mathbf{x_j}) - \mathbf{y_j})^2$$

ln 1/√(2πσ²)

(part in red is independent of w)

can be ignored

is exactly least squares estimate

① Ellipses in w

Optimization Trick

- Optimization Trick: Optimal point is invariant under monotonically increasing transformation (such as log) [Tutorial 2]
- $\log L(\mathbf{w}|\mathcal{D}) = LL(\mathbf{w}|\mathcal{D}) =$
  $$-\frac{m}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{m}(\mathbf{w^T}\phi(\mathbf{x_j}) - \mathbf{y_j})^2$$
  For a fixed (yet unknown) σ^2
  $$\hat{\mathbf{w}_{ML}} = \text{argmax}\ LL(y_1 \ldots y_m|\mathbf{x}_1 \ldots \mathbf{x}_m, \mathbf{w}, \sigma^2)$$
  $$= \text{argmin}_{w} \sum_{j=1}^{m}(\mathbf{w}^T\phi(\mathbf{x}_j) - y_j)^2$$

By flipping of sign on $-\frac{1}{2\sigma^2}\sum_j (w^T\phi(x_j) - y_j)^2$

Optimization Trick

- Optimization Trick: Optimal point is invariant under monotonically increasing transformation (such as log) [Tutorial 2]
- $\log L(\mathbf{w}|\mathcal{D}) = LL(\mathbf{w}|\mathcal{D}) =$
  $$-\frac{m}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{m}(\mathbf{w^T}\phi(\mathbf{x_j}) - \mathbf{y_j})^2$$
  For a fixed (yet unknown) σ^2
  $$\hat{\mathbf{w}_{ML}} = \text{argmax } LL(y_1 ... y_m|\mathbf{x}_1 \ldots \mathbf{x}_m, \mathbf{w}, \sigma^2)$$
  $$= \text{argmin } \sum_{j=1}^{m}(\mathbf{w}^T\phi(\mathbf{x}_j) - y_j)^2$$
- Note that this is same as the Least square solution!!

$$LL(\mathbf{w}|\mathcal{D}) = -\frac{m}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{m}(\mathbf{w^T}\phi(\mathbf{x_j}) - \mathbf{y_j})^2$$

$$LL(\mathbf{w}|\mathcal{D}) = \log\prod_{j=1}^{m}\Pr(\mathbf{w}|y_j, \mathbf{x}_j)$$

$$LL(\mathbf{w}|\mathcal{D}) = \prod_{j=1}^{m}\frac{1}{\sqrt{2\pi\sigma^2}}exp\left(\frac{(y_j - \mathbf{w}^T\phi(\mathbf{x}_j))^2}{2\sigma^2}\right)$$

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

Redundant Φ and Overfitting

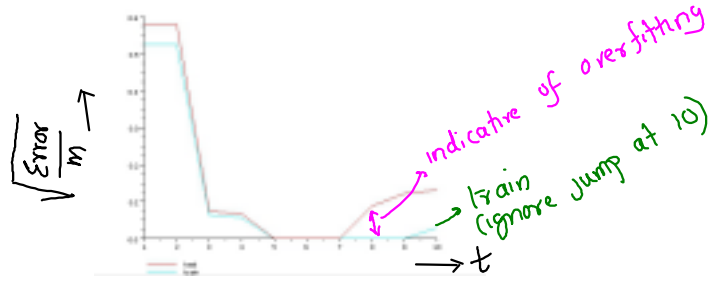
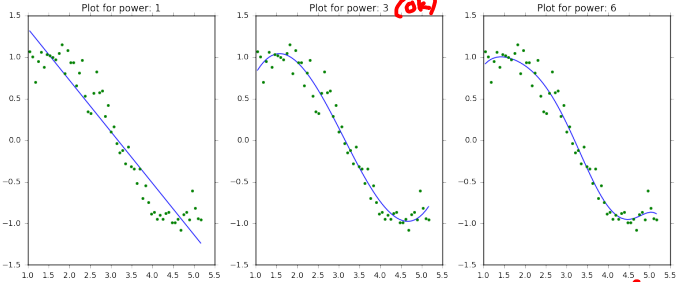
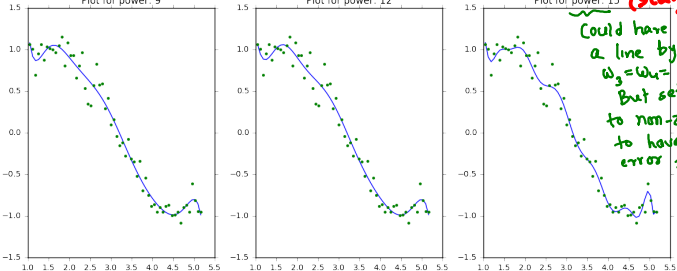


Figure 1: Root Mean Squared (RMS) errors on sample ~~train~~ and ~~test~~ datasets as a function of the degree t of the polynomial being fit

- Too many bends (t=9 onwards) in curve $\equiv$ high values of some $w_i's$. Try plotting values of w_i's using applet at `http://mste.illinois.edu/users/exner/java.f/leastsquares/#simulation`
- Train and test errors differ significantly

How does overfitting manifest itself? Pictorially





See Jupyter Notebooks

How does overfitting manifest itself? Weights

	rss	intercept	w1	w2	w3
model_pow_1	3.3	2	-0.62	NaN	NaN
model_pow_2	3.3	1.9	-0.58	-0.006	NaN
model_pow_3	1.1	-1.1	3	-1.3	0.14
model_pow_4	1.1	-0.27	1.7	-0.53	-0.036
model_pow_5	1	3	-5.1	4.7	-1.9
model_pow_6	0.99	-2.8	8.5	-9.7	5.2
model_pow_7	0.93	19	-56	69	-45
model_pow_8	0.92	43	-1.4e+02	1.8e+02	-1.3e+02
model_pow_9	0.87	1.7e+02	-6.1e+02	9.6e+02	-8.5e+02
model_pow_10	0.87	1.4e+02	-4.9e+02	7.3e+02	-8e+02
model_pow_11	0.87	-75	5.1e+02	-1.3e+03	1.9e+03
model_pow_12	0.87	-3.4e+02	1.9e+03	-4.4e+03	6e+03
model_pow_13	0.86	3.2e+03	-1.8e+04	4.5e+04	-6.7e+04
model_pow_14	0.79	2.4e+04	-1.4e+05	3.8e+05	-6.1e+05

$|w_i|$ increases (for each i) drastically as an indicator of overfitting (accounting for more & more curvature info)

We would like $|w_i|$ to be REGULATED!

See Jupyter Notebooks

Bayesian Linear Regression

- The Bayesian interpretation of probabilistic estimation is a logical extension that enables reasoning with uncertainty **but in the light of some background belief** ($|w_i|$ is bounded from above)
- **Bayesian linear regression**: A Bayesian alternative to **Maximum Likelihood** least squares regression
- Continue with Normally distributed errors
- Model the **w** using a prior distribution and use the posterior over **w** as the result
- Intuitive Prior: $w_i \sim N(0, \sigma_i^2)$ [Recall 3-σ rule]

Bayesian Linear Regression

- The Bayesian interpretation of probabilistic estimation is a logical extension that enables reasoning with uncertainty **but in the light of some background belief**
- **Bayesian linear regression**: A Bayesian alternative to **Maximum Likelihood** least squares regression
- Continue with Normally distributed errors
- Model the **w** using a prior distribution and use the posterior over **w** as the result
- Intuitive Prior: Components of **w** should not become too large!
- Next: Illustration of Bayesian Estimation on a simple Coin-tossing example

So far: Least Squares

- If solution $\Phi\mathbf{w} = \mathbf{y}$ exists, then least squares estimate $\mathbf{w}^*$ can be obtained by solving this linear system
- Additionally, if $n = m$ then Φ must be invertible and $\mathbf{w}^* = \Phi^{-1}\mathbf{y}$
- If $\mathbf{y}$ is NOT in the column space of Φ, then the least squares solution is obtained using the left-pseudoinverse of Φ:

$$\mathbf{w}^* = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y} \tag{1}$$

So far and next: MLE and Regularization

- The Maximum Likelihood estimate $\mathbf{w}_{MLE}$ happens to be the same as the least squares estimate $\mathbf{w}^*$. That is, $\mathbf{w}_{MLE} = \mathbf{w}^*$
- Here $\Phi^T\Phi$ is invertible only if Φ has full column rank
- Bayesian Estimation and Regularization: (a) Encode prior belief on $\mathbf{w}$ and (b) Develop probabilistic distributions on $\mathbf{w}^*$

Prior illustrated for simpler coin tossing & we will come back

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

Bayesian Linear Regression

- The Bayesian interpretation of probabilistic estimation is a logical extension that enables reasoning with uncertainty **but in the light of some background belief**
- **Bayesian linear regression**: A Bayesian alternative to **Maximum Likelihood** least squares regression
- Continue with Normally distributed errors
- Model the **w** using a prior distribution and use the posterior over **w** as the result
- Intuitive Prior:

Bayesian Linear Regression

- The Bayesian interpretation of probabilistic estimation is a logical extension that enables reasoning with uncertainty **but in the light of some background belief**
- **Bayesian linear regression**: A Bayesian alternative to **Maximum Likelihood** least squares regression
- Continue with Normally distributed errors
- Model the **w** using a prior distribution and use the posterior over **w** as the result
- Intuitive Prior: Components of **w** should not become too large!
- Next: Illustration of Bayesian Estimation on a simple Coin-tossing example

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

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?

Maximum Likelihood Estimation

- We are tempted to say that the probability of Heads in a subsequent toss is $400/1000 = 0.4$[1].
- But why?
- This is motivated by our wanting to maximize the probability of the data we have. Or in other words, we want a **Maximum Likelihood Estimate**.

[1] This raises an important point, you can never know the probability of the coin giving a head, what you can give is only an estimate for it. So dont be confused with 0.4 as the probability of getting a head, it is only an intelligent guess

Revisiting Likelihood

- Let the observed data follow a distribution f_θ, with θ being the unknown parameter.
  1. Coin tossing expt: θ is probability of heads occurring in any given toss and corresponds to a bernoulli distribution.
  2. Logistic regression: θ is basically w (a complex machine tossing coin)
- Let $X_1, X_2, ..., X_n$ be the set of random variables governing the observation with a joint pdf/pmf denoted by:

$$f_\theta(x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n|\theta)$$

Revisiting Likelihood: Continued

$x_i \in \{0, 1\}$

- The Maximum Likelihood Estimate (MLE) of θ is $\widehat{\theta}$ that maximises L(θ): $L(\theta) = f(x_1, x_2, ..., x_n|\theta)$
- For an independent and identically distributed sample $X_1, X_2, \ldots, X_n$, this means:

$$MLE(\theta) : \hat{\theta} = \underset{\theta}{\text{argmax}} \prod_{i=1}^{n} f(x_i|\theta)$$

$$\hat{\theta} = \underset{\theta}{\text{argmax}} \prod_{i=1}^{n} \theta^{x_i} * (1-\theta)^{(1-x_i)}$$

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.

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

($p \equiv \theta$)

If you did not know specific values of $x_1 \ldots x_{1000}$

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

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

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. Bernoulli & Binomial - Beta
2. Geometric - Beta
3. Categorical - Dirichlet
4. Multinomial - Dirichlet
5. Poisson - Gamma
6. Normal - Inverse Gamma

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 $Beta(p; \alpha, \beta) \sim \frac{p^{(\alpha-1)}(1-p)^{(\beta-1)}}{B(\alpha,\beta)}$,

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

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 $Beta(p; \alpha, \beta) \sim \frac{p^{(\alpha-1)}(1-p)^{(\beta-1)}}{B(\alpha,\beta)}$,

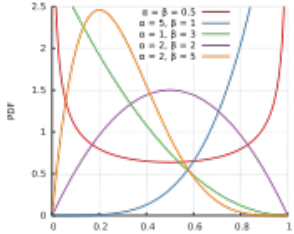
- $\Pr(H \mid \mathcal{D}) = \Pr(p \mid \mathcal{D}) = \dfrac{\Pr(\mathcal{D}|p)\Pr(p)}{\int_q \Pr(\mathcal{D} \mid q)\Pr(q)}$

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

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

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

1. $\mathbf{E}_{Beta(\alpha,\beta)}[p] = \frac{\alpha}{\alpha+\beta}$ and $\underset{p}{\text{argmax}}\, Beta(p; \alpha, \beta) = \dfrac{\alpha - 1}{\alpha + \beta - 2}$ (the mode of the distribution)
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:*
   $$\hat{p} = \underset{p}{\operatorname{argmax}} \ {}^{n}C_h p^h (1-p)^{n-h} = \frac{h}{n}$$
2. *The Maximum a-Posterior (MAP) Estimate:* The mode of the posterior distribution
   $$\tilde{p} = \underset{H}{\operatorname{argmax}} \Pr(H \mid \mathcal{D}) = \underset{p}{\operatorname{argmax}} \Pr(p \mid \mathcal{D})$$
   $$= \underset{p}{\operatorname{argmax}} \, Beta(p; \alpha + h, \beta + n - h) = \frac{\alpha + h - 1}{\alpha + \beta + n - 2}$$

Case Study Continued

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

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

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

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

Prior Distribution for w for Linear Regression

$$y = \mathbf{w}^T \phi(\mathbf{x}) + \varepsilon$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

- We saw that when we try to maximize log-likelihood we end up with $\hat{\mathbf{w}}_{MLE} = (\Phi^T \Phi)^{-1} \Phi^T y$
- We can use a Prior distribution on $\mathbf{w}$ to avoid over-fitting
  $$w_i \sim \mathcal{N}(0, \tfrac{1}{\lambda})$$
  (that is, each component w_i is approximately bounded within $\pm \frac{2}{\sqrt{\lambda}}$ by the $3 - \sigma$ rule)
- Q: How do deal with Bayesian Estimation for Gaussian distribution?

Conjugate Prior for (univariate) Gaussian

- We will temporarily generalize the discussion with x taking the place of ε and μ taking the place of w_i

Conjugate Prior for (univariate) Gaussian

- We will temporarily generalize the discussion with x taking the place of ε and μ taking the place of w_i
- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 ... x_m$
- $\mu_{MLE} = \frac{1}{m} \sum_{i=1}^{m} x_i$ and $\sigma_{MLE} = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)^2$
- The conjugate prior for the (univariate) normally distributed random variable X in the case that σ^2 is not a random variable is
  $\Pr(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$, And the **posterior** is?

Conjugate Prior for (univariate) Gaussian

- We will temporarily generalize the discussion with x taking the place of ε and μ taking the place of w_i
- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 ... x_m$
- $\mu_{MLE} = \frac{1}{m} \sum_{i=1}^{m} x_i$ and $\sigma_{MLE} = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)^2$
- The conjugate prior for the (univariate) normally distributed random variable X in the case that σ^2 is not a random variable is
  $\Pr(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$, And the **posterior** is?
- Answer: $\Pr(\mu | x_1 ... x_m) = \mathcal{N}(\mu_m, \sigma_m^2)$ such that

Conjugate Prior for (univariate) Gaussian

- We will temporarily generalize the discussion with x taking the place of ε and μ taking the place of w_i
- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 ... x_m$
- $\mu_{MLE} = \frac{1}{m} \sum_{i=1}^{m} x_i$ and $\sigma_{MLE} = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)^2$
- The conjugate prior for the (univariate) normally distributed random variable X in the case that σ^2 is not a random variable is
  $\Pr(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$, And the **posterior** is?
- Answer: $\Pr(\mu | x_1 ... x_m) = \mathcal{N}(\mu_m, \sigma_m^2)$ such that
- $\mu_m = (\frac{\sigma^2}{m\sigma_0^2 + \sigma^2} \mu_0) + (\frac{m\sigma_0^2}{m\sigma_0^2 + \sigma^2} \hat{\mu}_{ML})$