CS189/CS289A
Introduction to Machine Learning
Lecture 6:

Peter Bartlett

February 5, 2015

Outline

Outline

- Recall: Gaussian class conditionals lead to a logistic posterior.

Outline

- Recall: Gaussian class conditionals lead to a logistic posterior.
- Estimation

Outline

- Recall: Gaussian class conditionals lead to a logistic posterior.
- Estimation
  - Estimating the parameter of a Bernoulli random variable.

Outline

- Recall: Gaussian class conditionals lead to a logistic posterior.
- Estimation
  - Estimating the parameter of a Bernoulli random variable.
  - Estimating the parameters of a Gaussian random variable.

Outline

- Recall: Gaussian class conditionals lead to a logistic posterior.
- Estimation
  - Estimating the parameter of a Bernoulli random variable.
  - Estimating the parameters of a Gaussian random variable.
- Parameter estimation methods:

Outline

- Recall: Gaussian class conditionals lead to a logistic posterior.
- Estimation
  - Estimating the parameter of a Bernoulli random variable.
  - Estimating the parameters of a Gaussian random variable.
- Parameter estimation methods:
  - Method of moments.

Outline

- Recall: Gaussian class conditionals lead to a logistic posterior.
- Estimation
  - Estimating the parameter of a Bernoulli random variable.
  - Estimating the parameters of a Gaussian random variable.
- Parameter estimation methods:
  - Method of moments.
  - Maximum likelihood.

Outline

- Recall: Gaussian class conditionals lead to a logistic posterior.
- Estimation
  - Estimating the parameter of a Bernoulli random variable.
  - Estimating the parameters of a Gaussian random variable.
- Parameter estimation methods:
  - Method of moments.
  - Maximum likelihood.
  - Penalized maximum likelihood.

Outline

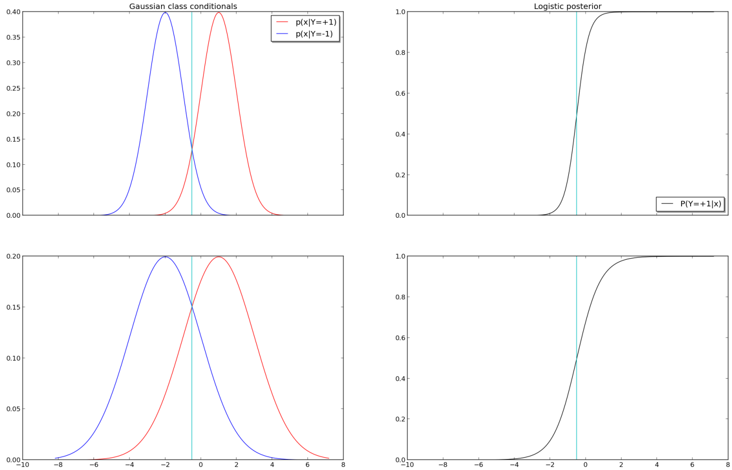
- Recall: Gaussian class conditionals lead to a logistic posterior.
- Estimation
  - Estimating the parameter of a Bernoulli random variable.
  - Estimating the parameters of a Gaussian random variable.
- Parameter estimation methods:
  - Method of moments.
  - Maximum likelihood.
  - Penalized maximum likelihood.
  - Bayesian estimates.

Outline

- **Recall: Gaussian class conditionals lead to a logistic posterior.**
- Estimation
  - Estimating the parameter of a Bernoulli random variable.
  - Estimating the parameters of a Gaussian random variable.
- Parameter estimation methods:
  - Method of moments.
  - Maximum likelihood.
  - Penalized maximum likelihood.
  - Bayesian estimates.

Gaussian generative to logistic discriminative models

$$p(x|y=+1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_+)^2}{2\sigma^2}\right).$$



Gaussian generative to logistic discriminative models

Class conditionals to posterior

For Gaussian class conditional densities with the same variance,

$$p(x|y = +1) = \mathcal{N}(\mu_+, \sigma^2), \qquad p(x|y = -1) = \mathcal{N}(\mu_-, \sigma^2),$$

the posterior probability is logistic

Gaussian generative to logistic discriminative models

Class conditionals to posterior

For Gaussian class conditional densities with the same variance,

$$p(x|y=+1) = \mathcal{N}(\mu_+, \sigma^2), \qquad p(x|y=-1) = \mathcal{N}(\mu_-, \sigma^2),$$

the posterior probability is logistic:

$$P(Y=+1|x) = \frac{1}{1+\exp\left(-x\cdot\theta-\theta_0\right)}.$$

Gaussian generative to logistic discriminative models

Class conditionals to posterior

For Gaussian class conditional densities with the same variance,

$$p(x|y=+1) = \mathcal{N}(\mu_+, \sigma^2), \qquad p(x|y=-1) = \mathcal{N}(\mu_-, \sigma^2),$$

the posterior probability is logistic:

$$P(Y=+1|x) = \frac{1}{1+\exp\left(-x\cdot\theta-\theta_0\right)}.$$

$$\theta = \frac{\mu_+ - \mu_-}{\sigma^2}, \qquad \theta_0 = \frac{\mu_-^2 - \mu_+^2}{2\sigma^2} - \log\left(\frac{P(-1)}{P(+1)}\right).$$

Outline

- Recall: Gaussian class conditionals lead to a logistic posterior.
- **Estimation.**
  - Estimating the parameter of a Bernoulli random variable.
  - Estimating the parameters of a Gaussian random variable.
- Parameter estimation methods:
  - Method of moments.
  - Maximum likelihood.
  - Penalized maximum likelihood.
  - Bayesian estimates.

Estimation

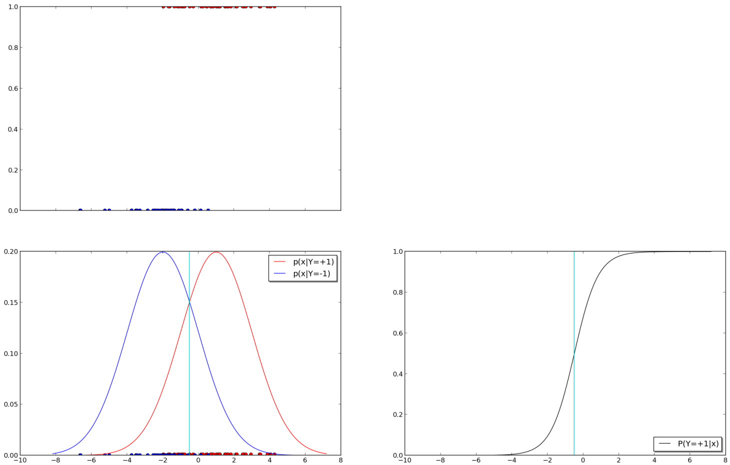
Estimation

- Suppose we want to use data to solve a classification problem.

Estimation

- Suppose we want to use data to solve a classification problem.
- How do we use the data to estimate the relevant probability distributions?

Estimating a Gaussian generative model



Estimation

- Suppose we want to use data to solve a classification problem.
- How do we use the data to estimate the relevant probability distributions?

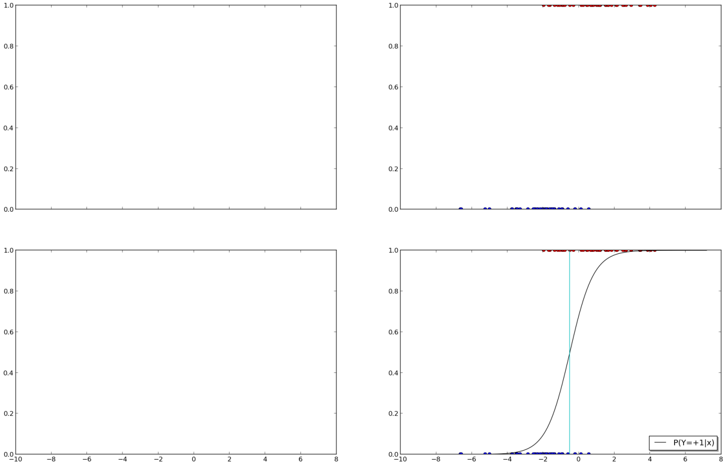
Estimation

- Suppose we want to use data to solve a classification problem.
- How do we use the data to estimate the relevant probability distributions?
- How do we estimate the class conditional distributions?

Estimation

- Suppose we want to use data to solve a classification problem.
- How do we use the data to estimate the relevant probability distributions?
- How do we estimate the class conditional distributions?
- How do we estimate the class probabilities?

Estimating a logistic discriminative model



Estimation

- Suppose we want to use data to solve a classification problem.
- How do we use the data to estimate the relevant probability distributions?
- How do we estimate the class conditional distributions?
- How do we estimate the class probabilities?

Estimation

- Suppose we want to use data to solve a classification problem.
- How do we use the data to estimate the relevant probability distributions?
- How do we estimate the class conditional distributions?
- How do we estimate the class probabilities?
- How do we estimate the posterior distribution?

Estimation

- Suppose we want to use data to solve a classification problem.
- How do we use the data to estimate the relevant probability distributions?
- How do we estimate the class conditional distributions?
- **How do we estimate the class probabilities?**
- How do we estimate the posterior distribution?

Outline

- Recall: Gaussian class conditionals lead to a logistic posterior.
- Estimation.
  - **Estimating the parameter of a Bernoulli random variable.**
  - Estimating the parameters of a Gaussian random variable.
- Parameter estimation methods:
  - Method of moments.
  - Maximum likelihood.
  - Penalized maximum likelihood.
  - Bayesian estimates.

Estimating a Bernoulli random variable

Estimating a Bernoulli random variable

Biased coin

Estimating a Bernoulli random variable

Biased coin

- We have a biased coin, $\Pr(+) = p$, $\Pr(-) = 1 - p$.

Estimating a Bernoulli random variable

Biased coin

- We have a biased coin, $\Pr(+) = p$, $\Pr(-) = 1 - p$.
- We don't know p.

Estimating a Bernoulli random variable

Biased coin

- We have a biased coin, $\Pr(+) = p$, $\Pr(-) = 1 - p$.
- We don't know p.
- We observe a sequence of outcomes:

$$+, +, -, -, +$$

Estimating a Bernoulli random variable

Biased coin

- We have a biased coin, $\Pr(+) = p$, $\Pr(-) = 1 - p$.
- We don't know p.
- We observe a sequence of outcomes:

$$+, +, -, -, +$$

- What is a good estimate of p?

Method of moments

Bernoulli estimate

Method of moments

Bernoulli estimate

- We could choose p so that the distribution it defines has the same expectation as the average of the data.
- The number of $+1$s we see from a single coin toss is a random variable with a *Bernoulli distribution*, $\Pr(1) = p$, $\Pr(0) = 1 - p$.
- We see n independent tosses.

Method of moments

Bernoulli estimate

- We could choose p so that the distribution it defines has the same expectation as the average of the data.
- The number of $+1$s we see from a single coin toss is a random variable with a *Bernoulli distribution*, $\Pr(1) = p$, $\Pr(0) = 1 - p$.
- We see n independent tosses. Define the number of $+1$s from each (either 0 or 1) as $X_1, X_2, \ldots, X_n$.

Method of moments

Bernoulli estimate

- We could choose p so that the distribution it defines has the same expectation as the average of the data.
- The number of $+1$s we see from a single coin toss is a random variable with a *Bernoulli distribution*, $\Pr(1) = p$, $\Pr(0) = 1 - p$.
- We see n independent tosses. Define the number of $+1$s from each (either 0 or 1) as $X_1, X_2, \ldots, X_n$. The average of these random variables is

$$\frac{1}{n}\sum_{i=1}^{n} X_i.$$

Method of moments

Bernoulli estimate

- We could choose p so that the distribution it defines has the same expectation as the average of the data.
- The number of $+1$s we see from a single coin toss is a random variable with a *Bernoulli distribution*, $\Pr(1) = p$, $\Pr(0) = 1 - p$.
- We see n independent tosses. Define the number of $+1$s from each (either 0 or 1) as $X_1, X_2, \ldots, X_n$. The average of these random variables is
$$\frac{1}{n} \sum_{i=1}^{n} X_i.$$
- To choose the parameter p of the distribution of the X_i so that the expectation is the average of the observed values, we choose $p =$.

Method of moments

Bernoulli estimate

- We could choose p so that the distribution it defines has the same expectation as the average of the data.
- The number of $+1$s we see from a single coin toss is a random variable with a *Bernoulli distribution*, $\Pr(1) = p$, $\Pr(0) = 1 - p$.
- We see n independent tosses. Define the number of $+1$s from each (either 0 or 1) as $X_1, X_2, \ldots, X_n$. The average of these random variables is

$$\frac{1}{n}\sum_{i=1}^{n} X_i.$$

- To choose the parameter p of the distribution of the X_i so that the expectation is the average of the observed values, we choose $p = 0.6$.

Maximum likelihood

Bernoulli estimate

Maximum likelihood

Bernoulli estimate

- We could choose p so that the distribution it defines gives the observed data the highest probability.

Maximum likelihood

Bernoulli estimate

- We could choose p so that the distribution it defines gives the observed data the highest probability.
- For a fixed choice of p,

$$\Pr(+,+,-,-,+) =$$

Maximum likelihood

Bernoulli estimate

- We could choose p so that the distribution it defines gives the observed data the highest probability.
- For a fixed choice of p,

$$\Pr(+,+,-,-,+) = p \cdot p \cdot (1-p) \cdot (1-p) \cdot p =$$

Maximum likelihood

Bernoulli estimate

- We could choose p so that the distribution it defines gives the observed data the highest probability.
- For a fixed choice of p,

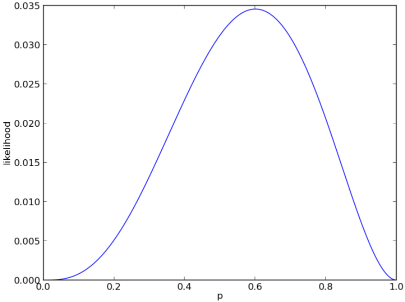
$$\Pr(+,+,-,-,+) = p \cdot p \cdot (1-p) \cdot (1-p) \cdot p = p^3(1-p)^2.$$

Maximum likelihood

Bernoulli estimate

- We could choose p so that the distribution it defines gives the observed data the highest probability.
- For a fixed choice of p,

$$\Pr(+,+,-,-,+) = p \cdot p \cdot (1-p) \cdot (1-p) \cdot p = p^3(1-p)^2.$$

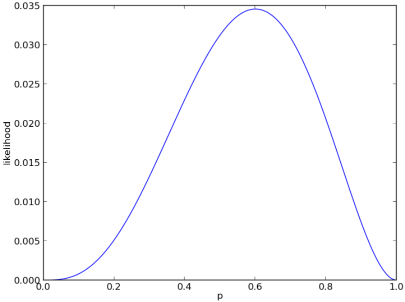


Maximum likelihood

Bernoulli estimate

- We could choose p so that the distribution it defines gives the observed data the highest probability.
- For a fixed choice of p,

$$\Pr(+,+,-,-,+) = p \cdot p \cdot (1-p) \cdot (1-p) \cdot p = p^3(1-p)^2.$$



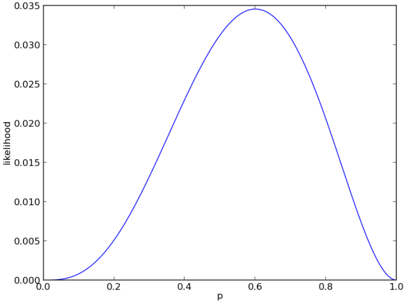
- The probability of the data under different choices of p, viewed as a function of p, is called the *likelihood*.

Maximum likelihood

Bernoulli estimate

- We could choose p so that the distribution it defines gives the observed data the highest probability.
- For a fixed choice of p,

$$\Pr(+,+,-,-,+) = p \cdot p \cdot (1-p) \cdot (1-p) \cdot p = p^3(1-p)^2.$$



- The probability of the data under different choices of p, viewed as a function of p, is called the *likelihood*.
- The maximizer of the likelihood in this case is $p = 0.6$.

Maximum likelihood

Bernoulli estimate

Maximum likelihood

Bernoulli estimate

- The method of moments and maximum likelihood give the same answer in this case.

Maximum likelihood

Bernoulli estimate

- The method of moments and maximum likelihood give the same answer in this case.
- In general, what is the maximum likelihood estimate for a Bernoulli?

Maximum likelihood

Bernoulli estimate

- The method of moments and maximum likelihood give the same answer in this case.
- In general, what is the maximum likelihood estimate for a Bernoulli?
- It's more convenient to work with the log likelihood, because the probability of a sequence of independent samples is a product of the probabilities, and the log of this product is a sum.

Maximum likelihood

Bernoulli estimate

- The method of moments and maximum likelihood give the same answer in this case.
- In general, what is the maximum likelihood estimate for a Bernoulli?
- It's more convenient to work with the log likelihood, because the probability of a sequence of independent samples is a product of the probabilities, and the log of this product is a sum.
- Maximizing log likelihood and maximizing likelihood are equivalent.

Maximum likelihood

Bernoulli estimate

- The method of moments and maximum likelihood give the same answer in this case.
- In general, what is the maximum likelihood estimate for a Bernoulli?
- It's more convenient to work with the log likelihood, because the probability of a sequence of independent samples is a product of the probabilities, and the log of this product is a sum.
- Maximizing log likelihood and maximizing likelihood are equivalent.

$$l(p) = \log \Pr(+,+,-,-,+) = \log\left(p^3(1-p)^2\right).$$

Maximum likelihood

Bernoulli estimate

- The method of moments and maximum likelihood give the same answer in this case.
- In general, what is the maximum likelihood estimate for a Bernoulli?
- It's more convenient to work with the log likelihood, because the probability of a sequence of independent samples is a product of the probabilities, and the log of this product is a sum.
- Maximizing log likelihood and maximizing likelihood are equivalent.

$$l(p) = \log \Pr(+,+,-,-,+) = \log\left(p^3(1-p)^2\right).$$

Set $0 = l'(p) = \frac{3}{p} - \frac{2}{1-p}$ and solve to get $p = 0.6$.

Maximum likelihood

Bernoulli estimate

- Suppose we saw k successes in n trials.

Maximum likelihood

Bernoulli estimate

- Suppose we saw k successes in n trials.
- The moment estimator gives $\hat{p} = \frac{k}{n}$.

Maximum likelihood

Bernoulli estimate

- Suppose we saw k successes in n trials.
- The moment estimator gives $\hat{p} = \frac{k}{n}$.
- What is the maximum likelihood estimate?

Maximum likelihood

Bernoulli estimate

- Suppose we saw k successes in n trials.
- The moment estimator gives $\hat{p} = \frac{k}{n}$.
- What is the maximum likelihood estimate?

$$\Pr(k \text{ of } n) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Maximum likelihood

Bernoulli estimate

- Suppose we saw k successes in n trials.
- The moment estimator gives $\hat{p} = \frac{k}{n}$.
- What is the maximum likelihood estimate?

$$\Pr(k \text{ of } n) = \binom{n}{k} p^k (1-p)^{n-k}.$$

$$l(p) = \log \Pr(k \text{ of } n) = \log \binom{n}{k} + k \log p + (n-k) \log(1-p).$$

Maximum likelihood

Bernoulli estimate

- Suppose we saw k successes in n trials.
- The moment estimator gives $\hat{p} = \frac{k}{n}$.
- What is the maximum likelihood estimate?

$$\Pr(k \text{ of } n) = \binom{n}{k} p^k (1-p)^{n-k}.$$

$$l(p) = \log \Pr(k \text{ of } n) = \log \binom{n}{k} + k \log p + (n-k) \log(1-p).$$

$$l'(p) = \frac{k}{p} - \frac{n-k}{1-p}.$$

Maximum likelihood

Bernoulli estimate

- Suppose we saw k successes in n trials.
- The moment estimator gives $\hat{p} = \frac{k}{n}$.
- What is the maximum likelihood estimate?

$$\Pr(k \text{ of } n) = \binom{n}{k} p^k (1-p)^{n-k}.$$

$$l(p) = \log \Pr(k \text{ of } n) = \log \binom{n}{k} + k \log p + (n-k) \log(1-p).$$

$$l'(p) = \frac{k}{p} - \frac{n-k}{1-p}.$$

$$\hat{p} = \frac{k}{n}.$$

Penalized maximum likelihood

Bernoulli estimate

- Suppose we think that p is close to $1/2$.

Penalized maximum likelihood

Bernoulli estimate

- Suppose we think that p is close to $1/2$.
- Then we see 0 successes in 3 trials.

Penalized maximum likelihood

Bernoulli estimate

- Suppose we think that p is close to $1/2$.
- Then we see 0 successes in 3 trials.
- The moment estimator and maximum likelihood give $\hat{p} = 0$.

Penalized maximum likelihood

Bernoulli estimate

- Suppose we think that p is close to $1/2$.
- Then we see 0 successes in 3 trials.
- The moment estimator and maximum likelihood give $\hat{p} = 0$.
- This might be unreasonable.

Penalized maximum likelihood

Bernoulli estimate

- Suppose we think that p is close to $1/2$.
- Then we see 0 successes in 3 trials.
- The moment estimator and maximum likelihood give $\hat{p} = 0$.
- This might be unreasonable.
- How can we incorporate our prior information that p is close to $1/2$ into our estimate?

Penalized maximum likelihood

Bernoulli estimate

- Suppose we think that p is close to $1/2$.
- Then we see 0 successes in 3 trials.
- The moment estimator and maximum likelihood give $\hat{p} = 0$.
- This might be unreasonable.
- How can we incorporate our prior information that p is close to $1/2$ into our estimate?
- We could use penalized maximum likelihood: maximize

Penalized maximum likelihood

Bernoulli estimate

- Suppose we think that p is close to $1/2$.
- Then we see 0 successes in 3 trials.
- The moment estimator and maximum likelihood give $\hat{p} = 0$.
- This might be unreasonable.
- How can we incorporate our prior information that p is close to $1/2$ into our estimate?
- We could use penalized maximum likelihood: maximize

$$\log(\Pr(0 \text{ of } 3)) + \log(p(1-p)) =$$

Penalized maximum likelihood

Bernoulli estimate

- Suppose we think that p is close to $1/2$.
- Then we see 0 successes in 3 trials.
- The moment estimator and maximum likelihood give $\hat{p} = 0$.
- This might be unreasonable.
- How can we incorporate our prior information that p is close to $1/2$ into our estimate?
- We could use penalized maximum likelihood: maximize

$$\log(\Pr(0 \text{ of } 3)) + \log(p(1-p)) = 3\log(1-p) + \log(p(1-p)).$$

Penalized maximum likelihood

Bernoulli estimate

- Suppose we think that p is close to $1/2$.
- Then we see 0 successes in 3 trials.
- The moment estimator and maximum likelihood give $\hat{p} = 0$.
- This might be unreasonable.
- How can we incorporate our prior information that p is close to $1/2$ into our estimate?
- We could use penalized maximum likelihood: maximize

$$\log(\Pr(0 \text{ of } 3)) + \log(p(1-p)) = 3\log(1-p) + \log(p(1-p)).$$

$$l'(p) = -\frac{3}{1-p} + \frac{1-2p}{p(1-p)};$$

Penalized maximum likelihood

Bernoulli estimate

- Suppose we think that p is close to $1/2$.
- Then we see 0 successes in 3 trials.
- The moment estimator and maximum likelihood give $\hat{p} = 0$.
- This might be unreasonable.
- How can we incorporate our prior information that p is close to $1/2$ into our estimate?
- We could use penalized maximum likelihood: maximize

$$\log(\Pr(0 \text{ of } 3)) + \log(p(1-p)) = 3\log(1-p) + \log(p(1-p)).$$

$$l'(p) = -\frac{3}{1-p} + \frac{1-2p}{p(1-p)}; \qquad \hat{p} = \frac{1}{5}.$$

Penalized maximum likelihood

Bernoulli estimate

- Suppose we think that p is close to $1/2$.
- Then we see 0 successes in 3 trials.
- The moment estimator and maximum likelihood give $\hat{p} = 0$.
- This might be unreasonable.
- How can we incorporate our prior information that p is close to $1/2$ into our estimate?
- We could use penalized maximum likelihood: maximize

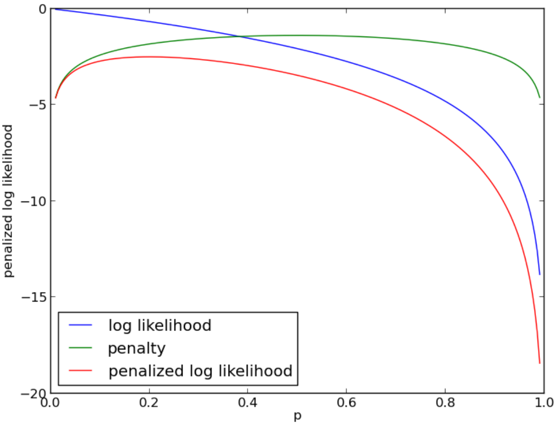
$$\log(\Pr(0 \text{ of } 3)) + \log(p(1-p)) = 3\log(1-p) + \log(p(1-p)).$$

$$l'(p) = -\frac{3}{1-p} + \frac{1-2p}{p(1-p)}; \qquad \hat{p} = \frac{1}{5}.$$

- Such estimators are particularly useful when the number of outcomes is not just 2 but large (for example, estimating the probability of words in a language).

Penalized maximum likelihood

0 successes in 3 trials:



Penalized maximum likelihood

$$\log(\Pr(k \text{ of } n)) + \log(p(1-p))$$

Penalized maximum likelihood

$$\begin{aligned}
&\log(\Pr(k \text{ of } n)) + \log(p(1-p)) \\
&\quad = \log\binom{n}{k} + k\log p + (n-k)\log(1-p) + \log(p(1-p)).
\end{aligned}$$

Penalized maximum likelihood

$$\log(\Pr(k \text{ of } n)) + \log(p(1-p))$$
$$= \log\binom{n}{k} + k\log p + (n-k)\log(1-p) + \log(p(1-p)).$$
$$l'(p) = \frac{k}{p} - \frac{n-k}{1-p} + \frac{1-2p}{p(1-p)}.$$

Penalized maximum likelihood

$$
\begin{aligned}
&\log(\Pr(k \text{ of } n)) + \log(p(1-p)) \\
&\qquad = \log\binom{n}{k} + k\log p + (n-k)\log(1-p) + \log(p(1-p)).
\end{aligned}
$$

$$
l'(p) = \frac{k}{p} - \frac{n-k}{1-p} + \frac{1-2p}{p(1-p)}.
$$

$$
p(n-k) = (1-p)k + 1 - 2p \qquad\qquad \hat{p} = \frac{k+1}{n+2}.
$$

Bayesian estimates

Bernoulli estimate

- Suppose we think that p is close to $1/2$.

Bayesian estimates

Bernoulli estimate

- Suppose we think that p is close to $1/2$.
- Another way to incorporate prior information about p:
  Rather than viewing p as an unknown number, model it as a random variable.

Bernoulli estimate

- Suppose we think that p is close to $1/2$.
- Another way to incorporate prior information about p:
  Rather than viewing p as an unknown number, model it as a random variable.
- Then our belief about the value of p is captured by a probability distribution over its possible values.

Bayesian estimates

Bernoulli estimate

- Suppose we think that p is close to $1/2$.
- Another way to incorporate prior information about p:
  Rather than viewing p as an unknown number, model it as a random variable.
- Then our belief about the value of p is captured by a probability distribution over its possible values.
- For example, if we have no *a priori* preference for one value of p over another, we might model p as a uniformly distributed random variable.

Bayesian estimates

Bernoulli estimate

- Suppose we think that p is close to $1/2$.
- Another way to incorporate prior information about p:
  Rather than viewing p as an unknown number, model it as a random variable.
- Then our belief about the value of p is captured by a probability distribution over its possible values.
- For example, if we have no *a priori* preference for one value of p over another, we might model p as a uniformly distributed random variable.
- Each observation allows us to update our belief, via Bayes Theorem.

Bayesian estimates

Update

Bayesian estimates

Update

prior distribution: $\pi(p) = 1$

Bayesian estimates

Update

$$\text{prior distribution:} \qquad \pi(p) = 1$$

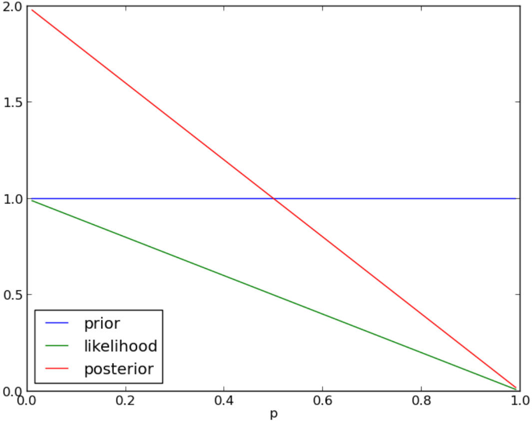
$$\text{posterior distribution:} \qquad P(p|X_1 = 1) \propto \underbrace{P(X_1 = 1|p)}_{\text{likelihood}} \underbrace{\pi(p)}_{\text{prior}}$$

Bayesian estimates

Update

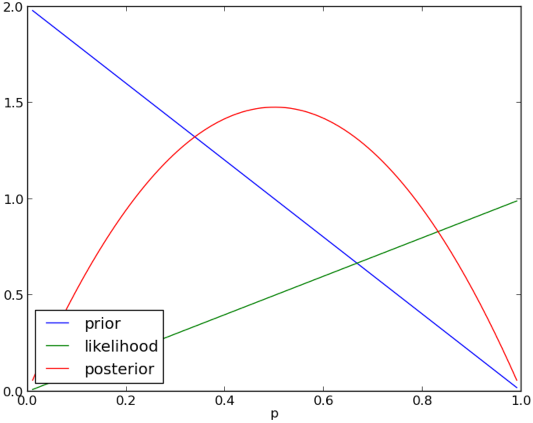
$$
\begin{aligned}
\text{prior distribution:} \quad & \pi(p) = 1 \\
\text{posterior distribution:} \quad & P(p|X_1 = 1) \propto \underbrace{P(X_1 = 1|p)}_{\text{likelihood}} \underbrace{\pi(p)}_{\text{prior}} \\
& = \frac{P(X_1 = 1|p)\pi(p)}{\int P(X_1 = 1|q)\, d\pi(q)}
\end{aligned}
$$

Prior $\pi(p) = 1$ on $[0, 1]$. $X_1 = 0$



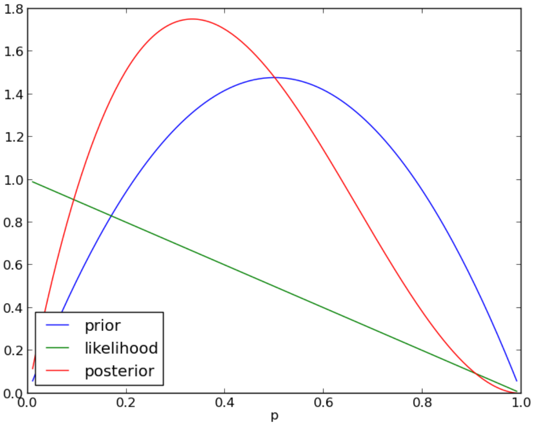
Bayesian estimates

$$X_2 = 1$$



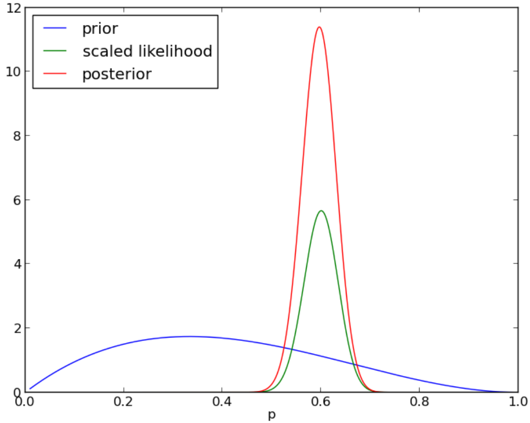
Bayesian estimates

$$X_3 = 0$$



Bayesian estimates

$$\frac{1}{200}\sum_{i=4}^{204} X_i = \frac{120}{200}$$



Bayesian estimates

Bernoulli estimation

- The posterior expresses our updated belief about the value of p.

Bayesian estimates

Bernoulli estimation

- The posterior expresses our updated belief about the value of p.
- We don't have a point estimate of p; we have a distribution over values that p might take.

Bayesian estimates

Bernoulli estimation

- The posterior expresses our updated belief about the value of p.
- We don't have a point estimate of p; we have a distribution over values that p might take.
- Notice that a Bayesian approach gives information about our uncertainty.

Bayesian estimates

Bernoulli estimation

- The posterior expresses our updated belief about the value of p.
- We don't have a point estimate of p; we have a distribution over values that p might take.
- Notice that a Bayesian approach gives information about our uncertainty.
- The Bayesian approach: assume that the parameters are randomly chosen with a fixed, known distribution.

Bayesian estimates

Bernoulli estimation

- The posterior expresses our updated belief about the value of p.
- We don't have a point estimate of p; we have a distribution over values that p might take.
- Notice that a Bayesian approach gives information about our uncertainty.
- The Bayesian approach: assume that the parameters are randomly chosen with a fixed, known distribution.
- Then everything in our prediction problem is a random variable with a known distribution. In that sense, there are no unknowns, just unobserved random variables with known distributions.

Bayesian estimates

Bernoulli estimation

- The posterior expresses our updated belief about the value of p.
- We don't have a point estimate of p; we have a distribution over values that p might take.
- Notice that a Bayesian approach gives information about our uncertainty.
- The Bayesian approach: assume that the parameters are randomly chosen with a fixed, known distribution.
- Then everything in our prediction problem is a random variable with a known distribution. In that sense, there are no unknowns, just unobserved random variables with known distributions.
- Bayesian estimation is just a computation: compute a conditional probability distribution.

Bayesian estimates

Bernoulli estimation

- If we need a point estimate, we might use the MAP estimate (maximum a posteriori probability): the mode of the posterior.

Bayesian estimates

Bernoulli estimation

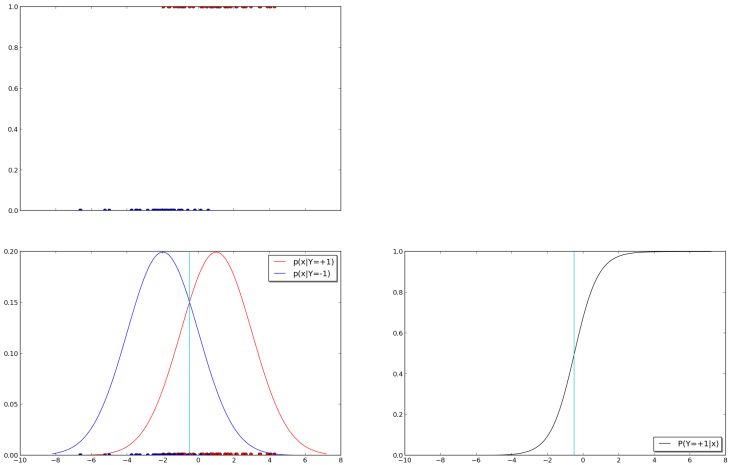
- If we need a point estimate, we might use the MAP estimate (maximum a posteriori probability): the mode of the posterior.
- The MAP estimate with a uniform prior corresponds to the maximum likelihood estimate.

Bayesian estimates

Bernoulli estimation

- If we need a point estimate, we might use the MAP estimate (maximum a posteriori probability): the mode of the posterior.
- The MAP estimate with a uniform prior corresponds to the maximum likelihood estimate.
- The MAP estimate with any other prior corresponds to a penalized maximum likelihood estimate. For instance, the penalty we considered earlier corresponds to a prior proportional to $p(1-p)$

Estimating a Gaussian generative model



Estimation

- Suppose we want to use data to solve a classification problem.
- How do we use the data to estimate the relevant probability distributions?
- How do we estimate the class probabilities? **Estimate a Bernoulli.**
- How do we estimate the class conditional distributions?

Estimation

- Suppose we want to use data to solve a classification problem.
- How do we use the data to estimate the relevant probability distributions?
- How do we estimate the class probabilities? Estimate a Bernoulli.
- **How do we estimate the class conditional distributions?**

Estimation

- Estimating the class probabilities corresponds to estimating a single parameter.

Estimation

- Estimating the class probabilities corresponds to estimating a single parameter.
- Class conditional distributions are much richer.

Estimation

- Estimating the class probabilities corresponds to estimating a single parameter.
- Class conditional distributions are much richer.
- To estimate a class conditional distribution, one approach is to assume that the distribution comes from a parameterized family, and estimate the parameters.

Estimation

- Estimating the class probabilities corresponds to estimating a single parameter.
- Class conditional distributions are much richer.
- To estimate a class conditional distribution, one approach is to assume that the distribution comes from a parameterized family, and estimate the parameters.
- For instance, it might be reasonable to assume that the class conditional distribution is a Gaussian.

Estimation

- Estimating the class probabilities corresponds to estimating a single parameter.
- Class conditional distributions are much richer.
- To estimate a class conditional distribution, one approach is to assume that the distribution comes from a parameterized family, and estimate the parameters.
- For instance, it might be reasonable to assume that the class conditional distribution is a Gaussian.
- Then we need to estimate the mean and variance.

Estimation

- Estimating the class probabilities corresponds to estimating a single parameter.
- Class conditional distributions are much richer.
- To estimate a class conditional distribution, one approach is to assume that the distribution comes from a parameterized family, and estimate the parameters.
- For instance, it might be reasonable to assume that the class conditional distribution is a Gaussian.
- Then we need to estimate the mean and variance.
- How can we do that?

Outline

- Recall: Gaussian class conditionals lead to a logistic posterior.
- Estimation.
  - Estimating the parameter of a Bernoulli random variable.
  - **Estimating the parameters of a Gaussian random variable.**
- Parameter estimation methods:
  - Method of moments.
  - Maximum likelihood.
  - Penalized maximum likelihood.
  - Bayesian estimates.

Gaussian Estimation

- We have a Gaussian distributed random variable

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Gaussian Estimation

- We have a Gaussian distributed random variable

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

- We don't know μ, σ^2.

Gaussian Estimation

- We have a Gaussian distributed random variable

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

- We don't know μ, σ^2.
- We observe a sequence of outcomes:

$$0.470, 3.346, -0.898, 2.155, -0.092$$

Gaussian Estimation

- We have a Gaussian distributed random variable

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

- We don't know μ, σ^2.
- We observe a sequence of outcomes:

$$0.470, 3.346, -0.898, 2.155, -0.092$$

- What is a good estimate of $\theta = (\mu, \sigma^2)$?

Method of moments

Gaussian Estimation

- We could choose $\theta = (\mu, \sigma^2)$ so that the distribution it defines has the same expectation as the average of the data.

Method of moments

Gaussian Estimation

- We could choose $\theta = (\mu, \sigma^2)$ so that the distribution it defines has the same expectation as the average of the data.
- The expectation of a Gaussian with parameters (μ, σ^2) is μ.

Method of moments

Gaussian Estimation

- We could choose $\theta = (\mu, \sigma^2)$ so that the distribution it defines has the same expectation as the average of the data.
- The expectation of a Gaussian with parameters (μ, σ^2) is μ.
- To choose the parameter of the distribution of the Gaussian so that the expectation is the average of the observed values, we choose

$$\mu = \frac{1}{n} \sum_{i=1}^{n} X_i = 0.996.$$

Method of moments

Gaussian Estimation

- We could choose $\theta = (\mu, \sigma^2)$ so that the distribution it defines has the same expectation as the average of the data.
- The expectation of a Gaussian with parameters (μ, σ^2) is μ.
- To choose the parameter of the distribution of the Gaussian so that the expectation is the average of the observed values, we choose

$$\mu = \frac{1}{n} \sum_{i=1}^{n} X_i = 0.996.$$

- What about σ^2?

Method of moments

Gaussian Estimation

- We could choose $\theta = (\mu, \sigma^2)$ so that the distribution it defines has the same first moment ($\mathbb{E}X$) *and second moment* ($\mathbb{E}X^2$) as the data.

Method of moments

Gaussian Estimation

- We could choose $\theta = (\mu, \sigma^2)$ so that the distribution it defines has the same first moment ($\mathbb{E}X$) *and second moment* ($\mathbb{E}X^2$) as the data.
- The variance of a Gaussian with parameters (μ, σ^2) is σ^2. So the second moment is $\mathbb{E}X^2 = \mathbb{E}(X - \mu)^2 + \mu^2 = \sigma^2 + \mu^2$.

Method of moments

Gaussian Estimation

- We could choose $\theta = (\mu, \sigma^2)$ so that the distribution it defines has the same first moment ($\mathbb{E}X$) *and second moment* ($\mathbb{E}X^2$) as the data.
- The variance of a Gaussian with parameters (μ, σ^2) is σ^2. So the second moment is $\mathbb{E}X^2 = \mathbb{E}(X - \mu)^2 + \mu^2 = \sigma^2 + \mu^2$.
- To match first and second moments, we choose

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mu^2 = 2.38.$$

Method of moments

Gaussian Estimation

- We could choose $\theta = (\mu, \sigma^2)$ so that the distribution it defines has the same first moment ($\mathbb{E}X$) *and second moment* ($\mathbb{E}X^2$) as the data.
- The variance of a Gaussian with parameters (μ, σ^2) is σ^2. So the second moment is $\mathbb{E}X^2 = \mathbb{E}(X - \mu)^2 + \mu^2 = \sigma^2 + \mu^2$.
- To match first and second moments, we choose

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mu^2 = 2.38.$$

- In general, if we have p parameters, we can solve p equations, and so need to match the corresponding number of moments.

Maximum likelihood

Gaussian Estimation

Maximum likelihood

Gaussian Estimation

- We could choose $\theta = (\mu, \sigma^2)$ so that the distribution it defines gives the observed data the highest probability density.

Maximum likelihood

Gaussian Estimation

- We could choose $\theta = (\mu, \sigma^2)$ so that the distribution it defines gives the observed data the highest probability density.
- For a fixed choice of (μ, σ^2),

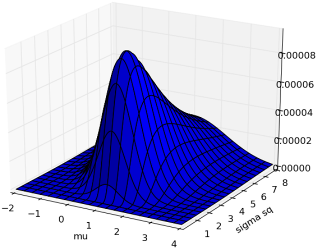
$$p(x_1)p(x_2)\cdots p(x_5) = \frac{1}{(2\pi\sigma^2)^{5/2}} \exp\left(-\frac{\sum_{i=1}^{5}(x_i - \mu)^2}{2\sigma^2}\right).$$

Maximum likelihood

Gaussian Estimation

- We could choose $\theta = (\mu, \sigma^2)$ so that the distribution it defines gives the observed data the highest probability density.
- For a fixed choice of (μ, σ^2),

$$p(x_1)p(x_2)\cdots p(x_5) = \frac{1}{(2\pi\sigma^2)^{5/2}} \exp\left(-\frac{\sum_{i=1}^{5}(x_i - \mu)^2}{2\sigma^2}\right).$$

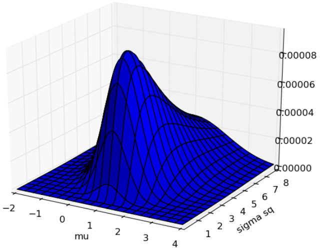


Maximum likelihood

Gaussian Estimation

- We could choose $\theta = (\mu, \sigma^2)$ so that the distribution it defines gives the observed data the highest probability density.
- For a fixed choice of (μ, σ^2),

$$p(x_1)p(x_2)\cdots p(x_5) = \frac{1}{(2\pi\sigma^2)^{5/2}} \exp\left(-\frac{\sum_{i=1}^{5}(x_i - \mu)^2}{2\sigma^2}\right).$$



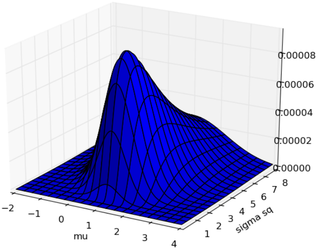
- The likelihood is the density of the data under different choices of $\theta = (\mu, \sigma^2)$, viewed as a function of θ.

Maximum likelihood

Gaussian Estimation

- We could choose $\theta = (\mu, \sigma^2)$ so that the distribution it defines gives the observed data the highest probability density.
- For a fixed choice of (μ, σ^2),

$$p(x_1)p(x_2)\cdots p(x_5) = \frac{1}{(2\pi\sigma^2)^{5/2}} \exp\left(-\frac{\sum_{i=1}^{5}(x_i-\mu)^2}{2\sigma^2}\right).$$



- The likelihood is the density of the data under different choices of $\theta = (\mu, \sigma^2)$, viewed as a function of θ.
- The maximizer of the likelihood is $\mu = 0.996$, $\sigma^2 = 2.38$.

Maximum likelihood

Gaussian Estimation

- The method of moments and maximum likelihood again give the same answer.

Maximum likelihood

Gaussian Estimation

- The method of moments and maximum likelihood again give the same answer.
- In general, what is the maximum likelihood estimate for a Gaussian?

Maximum likelihood

Gaussian Estimation

- The method of moments and maximum likelihood again give the same answer.
- In general, what is the maximum likelihood estimate for a Gaussian?
- Again, it's more convenient to work with the log likelihood.

$$l(\mu, \sigma^2) = \log \left(\prod_{i=1}^{n} \left(\frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right) \right)$$

Maximum likelihood

Gaussian Estimation

- The method of moments and maximum likelihood again give the same answer.
- In general, what is the maximum likelihood estimate for a Gaussian?
- Again, it's more convenient to work with the log likelihood.

$$
\begin{aligned}
l(\mu, \sigma^2) &= \log \left(\prod_{i=1}^{n} \left(\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right) \right) \\
&= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.
\end{aligned}
$$

Maximum likelihood

Gaussian Estimation

- The method of moments and maximum likelihood again give the same answer.
- In general, what is the maximum likelihood estimate for a Gaussian?
- Again, it's more convenient to work with the log likelihood.

$$
\begin{aligned}
l(\mu, \sigma^2) &= \log \left(\prod_{i=1}^{n} \left(\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right) \right) \\
&= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 .
\end{aligned}
$$

Set $\nabla l(\theta) = 0$ and solve.

Maximum likelihood

Gaussian maximum likelihood estimation

$\nabla l(\theta) = 0$ for

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \frac{n}{2\sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Other estimators for Gaussian parameters

Gaussian Estimation

Other estimators for Gaussian parameters

Gaussian Estimation

- We can also use penalized maximum likelihood estimators:

$$
\begin{aligned}
& l(\mu, \sigma^2) - \text{penalty}(\mu, \sigma^2) \\
& = \log \left(\prod_{i=1}^{n} \left(\frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right) \right) - \text{penalty}(\mu, \sigma^2)
\end{aligned}
$$

Other estimators for Gaussian parameters

Gaussian Estimation

- We can also use penalized maximum likelihood estimators:

$$
\begin{aligned}
&l(\mu, \sigma^2) - \text{penalty}(\mu, \sigma^2) \\
&= \log \left(\prod_{i=1}^{n} \left(\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right) \right) - \text{penalty}(\mu, \sigma^2)
\end{aligned}
$$

- And Bayesian estimators:

$$
\begin{aligned}
\text{prior distribution:} \quad & \pi(\theta) = 1 \\
\text{posterior distribution:} \quad & p(\theta | X_1 = x_1) \propto \underbrace{p(X_1 = x_1 | \theta)}_{\text{likelihood}} \underbrace{\pi(\theta)}_{\text{prior}} \\
& = \frac{p(X_1 = x_1 | \theta) \pi(\theta)}{\int p(X_1 = x_1 | q) \, d\pi(q)}
\end{aligned}
$$

Other estimators for Gaussian parameters

Gaussian Estimation

- Penalized maximum likelihood estimators and Bayesian estimators are particularly effective in the high-dimensional setting, when the number of parameters is large compared to the amount of data.

Outline

- Recall: Gaussian class conditionals lead to a logistic posterior.
- Estimation.
    - Estimating the parameter of a Bernoulli random variable.
    - Estimating the parameters of a Gaussian random variable.
- Parameter estimation methods:
    - Method of moments.
    - Maximum likelihood.
    - Penalized maximum likelihood.
    - Bayesian estimates.