Introduction to Bayesian Learning

Machine Learning
Fall 2023

The slides are partly from Vivek Srikumar
What we have seen so far

What does it mean to learn?

- Mistake-driven learning
  - Learning by counting (and bounding) number of mistakes
- PAC learnability
  - Sample complexity and bounds on errors on unseen examples

Various learning algorithms

- Analyzed algorithms under these models of learnability
- In all cases, the algorithm outputs a function that produces a label $y$ for a given input $x$
Coming up

Another way of thinking about “What does it mean to learn?”

- Bayesian learning

Different learning algorithms in this regime

- Logistic Regression
- Naïve Bayes
Today’s lecture

• Bayesian Learning

• Maximum a posteriori (MAP) and maximum likelihood (ML) estimation

• Two examples of maximum likelihood (ML) estimation
  – Binomial distribution
  – Normal distribution
Today’s lecture

• Bayesian Learning

• Maximum a posteriori (MAP) and maximum likelihood (ML) estimation

• Two examples of maximum likelihood estimation
  – Binomial distribution
  – Normal distribution
Probabilistic Learning

Two different notions of probabilistic learning
Probabilistic Learning

Two different notions of probabilistic learning

• **Learning probabilistic concepts**
  – The learned concept is a function $c : X \rightarrow [0,1]$
  – $c(x)$ may be interpreted as the probability that the label 1 is assigned to $x$
  – The learning theory that we have studied before is applicable (with some extensions)
Probabilistic Learning

Two different notions of probabilistic learning

• **Learning probabilistic concepts**
  – The learned concept is a function $c : X \rightarrow [0,1]$
  – $c(x)$ may be interpreted as the probability that the label 1 is assigned to $x$
  – The learning theory that we have studied before is applicable (with some extensions)

• **Bayesian Learning**: Use of a probabilistic criterion in selecting a hypothesis
  – The hypothesis can be deterministic, e.g., a Boolean function
  – The criterion for selecting the hypothesis is probabilistic
Bayesian Learning: The basics

- **Goal**: To find the **best** hypothesis from some space $H$ of hypotheses, using the observed data $D$

- Define **best** = most probable hypothesis in $H$
Bayesian Learning: The basics

- **Goal**: To find the *best* hypothesis from some space $H$ of hypotheses, using the observed data $D$

- Define *best* = *most probable hypothesis* in $H$

- In order to do that, we need to assume a probability distribution *over the class* $H$
Bayesian Learning: The basics

• **Goal:** To find the *best* hypothesis from some space $H$ of hypotheses, using the observed data $D$

• Define *best* = *most probable hypothesis* in $H$

• In order to do that, we need to assume a probability distribution *over the class* $H$

• We also need to know something about the relation between the data observed and the hypotheses
Bayesian methods have multiple benefits

• Provide interpretable learning algorithms

• Combining prior knowledge with observed data
  – Guide the model towards something we know

• Provide tools for learning and analyzing learning
  – Integrate uncertainty of the learning results
Bayes Theorem

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \]
Bayes Theorem

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \]

∀\(x, y\) \[ P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)} \]
Bayes Theorem

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \]
Bayes Theorem

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \]

**Prior probability**: What is our belief in Y before we see X?
Bayes Theorem

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \]

- **Likelihood**: What is the likelihood of observing $X$ given a specific $Y$?
- **Prior probability**: What is our belief in $Y$ before we see $X$?
Bayes Theorem

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \]

*Posterior probability*: What is the probability of Y given that X is observed?

*Likelihood*: What is the likelihood of observing X given a specific Y?

*Prior probability*: What is our belief in Y before we see X?
Bayes Theorem

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \]

- **Posterior probability**: What is the probability of Y given that X is observed?
- **Likelihood**: What is the likelihood of observing X given a specific Y?
- **Prior probability**: What is our belief in Y before we see X?

**Posterior \propto Likelihood \times Prior**
Probability Refresher

• **Product rule:** \( P(A \land B) = P(A, B) = P(A|B)P(B) = P(B|A)P(A) \)

• **Sum rule:** \( P(A \lor B) = P(A) + P(B) - P(A, B) \)

• Events A, B are **independent** if:
  
  - \( P(A, B) = P(A) P(B) \)
  
  - Equivalently, \( P(A \mid B) = P(A) \), \( P(B \mid A) = P(B) \)

• **Theorem of Total probability:**

  For mutually exclusive events \( A_1, A_2, \ldots, A_n \) (i.e. \( A_i \cap A_j = \emptyset \)) with \( \sum_i P(A_i) = 1 \)

  \[
P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)
  \]
Bayesian Learning

Given a dataset D, we want to find the **best** hypothesis h

What does *best* mean?

Bayesian learning uses $P(h \mid D)$, the conditional probability of a hypothesis given the data, to define *best*. 
Bayesian Learning

Given a dataset $D$, we want to find the best hypothesis $h$
What does *best* mean?

$$P(h|D)$$
Bayesian Learning

Given a dataset $D$, we want to find the best hypothesis $h$

What does *best* mean?

$$P(h|D)$$

*Posterior probability*: What is the probability that $h$ is the true hypothesis, given that the data $D$ is observed?
Bayesian Learning

Given a dataset D, we want to find the best hypothesis h.

What does *best* mean?

\[ P(h|D) \]

*Posterior probability*: What is the probability that h is the true hypothesis, given that the data D is observed?

*Key insight*: Both h and D are events.

- D: The event that we observed *this* particular dataset
- h: The event that the hypothesis h is the true hypothesis

So we can apply the Bayes rule here.
Bayesian Learning

Given a dataset $D$, we want to find the best hypothesis $h$. What does *best* mean?

Key insight: Both $h$ and $D$ are events.
- $D$: The event that we observed *this* particular dataset
- $h$: The event that the hypothesis $h$ is the true hypothesis

Posterior probability: What is the probability that $h$ is the true hypothesis, given that the data $D$ is observed?

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$
Bayesian Learning

Given a dataset $D$, we want to find the best hypothesis $h$.
What does best mean?

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

*Posterior probability:* What is the probability that $h$ is the true hypothesis, given that the data $D$ is observed?

*Prior probability of $h$:*
Background knowledge. What do we expect the true hypothesis to be even before we see any data? For example, in the absence of any information, maybe the uniform distribution.
Bayesian Learning

Given a dataset D, we want to find the best hypothesis h. What does best mean?

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

**Posterior probability**: What is the probability that h is the true hypothesis, given that the data D is observed?

**Likelihood**: What is the probability that this data point (an example or an entire dataset) is observed, given that the true hypothesis is h?

**Prior probability of h**: Background knowledge. What do we expect the true hypothesis to be even before we see any data? For example, in the absence of any information, maybe the uniform distribution.
Bayesian Learning

Given a dataset $D$, we want to find the best hypothesis $h$.

What does *best* mean?

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

*Posterior probability*: What is the probability that $h$ is the true hypothesis, given that the data $D$ is observed?

*Likelihood*: What is the probability that this data point (an example or an entire dataset) is observed, given that the hypothesis is $h$?

*Prior probability of $h$*: Background knowledge. What do we expect the true hypothesis to be even before we see any data? For example, in the absence of any information, maybe the uniform distribution.

What is the probability that the data $D$ is observed (independent of any knowledge about the hypothesis)?
Outline

• Bayesian Learning

• Maximum a posteriori and maximum likelihood estimation

• Two examples of maximum likelihood estimation
  – Binomial distribution
  – Normal distribution
Choosing a hypothesis

Given some data, find the most probable hypothesis

- The Maximum a Posteriori (MAP) hypothesis $h_{\text{MAP}}$

$$h_{\text{MAP}} = \arg \max_{h \in H} P(h|D)$$
Choosing a hypothesis

Given some data, find the most probable hypothesis

- The **Maximum a Posteriori (MAP)** hypothesis $h_{MAP}$

$$h_{MAP} = \arg \max_{h \in H} P(h|D)$$

$$= \arg \max_{h \in H} \frac{P(D|h)P(h)}{P(D)}$$

$$= \arg \max_{h \in H} P(D|h)P(h)$$
Choosing a hypothesis

Given some data, find the most probable hypothesis

– The Maximum a Posteriori (MAP) hypothesis $h_{\text{MAP}}$

$$h_{\text{MAP}} = \arg\max_{h \in H} P(h|D)$$

$$= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)}$$

$$= \arg\max_{h \in H} P(D|h)P(h)$$

Posterior $\propto$ Likelihood $\times$ Prior
Choosing a hypothesis

Given some data, find the most probable hypothesis

- The **Maximum a Posteriori (MAP)** hypothesis \( h_{MAP} \)

\[
h_{MAP} = \arg \max_{h \in H} P(D|h)P(h)
\]
Choosing a hypothesis

Given some data, find the most probable hypothesis

- The **Maximum a Posteriori (MAP)** hypothesis $h_{MAP}$

$$h_{MAP} = \arg \max_{h \in H} P(D|h)P(h)$$

If we assume that the prior is uniform i.e. $P(h_i) = P(h_j)$, for all $h_i, h_j$

- Simplify this to get the **Maximum Likelihood (ML)** hypothesis

$$h_{ML} = \arg \max_{h \in H} P(D|h)$$
Choosing a hypothesis

Given some data, find the most probable hypothesis

- The **Maximum a Posteriori (MAP)** hypothesis $h_{MAP}$

$$h_{MAP} = \arg \max_{h \in H} P(D|h)P(h)$$

If we assume that the prior is uniform (i.e. $P(h_i) = P(h_j)$, for all $h_i, h_j$)

- Simplify this to get the **Maximum Likelihood (ML)** hypothesis

$$h_{ML} = \arg \max_{h \in H} P(D|h)$$

Often computationally easier to maximize log likelihood
Today’s lecture

• Bayesian Learning

• Maximum a posteriori and maximum likelihood estimation

• Two examples of maximum likelihood estimation
  – Bernoulli trials
  – Normal distribution
Maximum Likelihood estimation

Maximum Likelihood estimation (MLE)

\[ h_{ML} = \arg \max_{h \in H} P(D|h) \]

What we need in order to define learning:

1. A hypothesis space H
2. A model that says how data D is generated given h
Example 1: Bernoulli trials

The CEO of a startup hires you for your first consulting job

- **CEO:** My company makes light bulbs. I need to know what is the probability they are faulty.

- **You:** Sure. I can help you out. Are they all identical?

- **CEO:** Yes!

- **You:** Excellent. I know how to help. We need to experiment...
Faulty lightbulbs

The experiment:
Try out 100 lightbulbs
80 work, 20 don’t

You: The probability is $P(\text{failure}) = 0.2$

CEO: But how do you know?

You: Because...
Bernoulli trials

- \( P(\text{failure}) = p, \ P(\text{success}) = 1 - p \)

- Each trial is i.i.d
  - Independent and identically distributed
Bernoulli trials

- $P(\text{failure}) = p$, $P(\text{success}) = 1 - p$

- Each trial is i.i.d
  - Independent and identically distributed

- You have seen $D = \{80 \text{ work, 20 don’t}\}$

$$P(D|p) = \binom{100}{80} p^{80} (1 - p)^{20}$$
Bernoulli trials

• P(failure) = p, P(success) = 1 − p

• Each trial is i.i.d
  – Independent and identically distributed

• You have seen D = {80 work, 20 don’t}

\[ P(D|p) = \binom{100}{80} p^{80} (1 - p)^{20} \]

• The most likely value of p for this observation is?
Bernoulli trials

- $P(\text{failure}) = p$, $P(\text{success}) = 1 - p$
- Each trial is i.i.d
  - Independent and identically distributed
- You have seen $D = \{80 \text{ work, 20 don’t}\}$
  \[
P(D|p) = \binom{100}{80} p^{80} (1 - p)^{20}
\]
- The most likely value of $p$ for this observation is?
  \[
  \arg\max_p P(D|p) = \arg\max_p \binom{100}{80} p^{80} (1 - p)^{20}
  \]
The “learning” algorithm

Say you have \( a \) Work and \( b \) Not-Work

\[
p_{best} = \operatorname{argmax}_p P(D|h)
\]
The “learning” algorithm

Say you have $a$ Work and $b$ Not-Work

\[
\rho_{\text{best}} = \arg \max_p P(D|h)
\]

\[
= \arg \max_p \log P(D|h)
\]

\[
= \arg \max_p a \log p + b \log (1 - p)
\]

Log likelihood
The “learning” algorithm

Say you have a Work and b Not-Work

\[ p_{best} = \arg \max_p P(D|h) \]

\[ = \arg \max_p \log P(D|h) \]

\[ = \arg \max_p \log \left( \frac{a + b}{a} p^a (1 - p)^b \right) \]

\[ = \arg \max_p a \log p + b \log (1 - p) \]
The “learning” algorithm

Say you have a Work and b Not-Work

\[ p_{best} = \arg\max_p P(D|h) \]

\[ = \arg\max_p \log P(D|h) \]

\[ = \arg\max_p \log \left( \left( \frac{a+b}{a} \right)^p \frac{a}{p} \frac{1-p}{1-p} \right) \]

\[ = \arg\max_p a \log p + b \log(1-p) \]

Calculus 101: Set the derivative to zero

Log likelihood
The “learning” algorithm

Say you have a Work and b Not-Work

\[ p_{best} = \arg\max_p P(D|h) \]
\[ = \arg\max_p \log P(D|h) \]
\[ = \arg\max_p \log \left( \frac{a + b}{a} p^a (1 - p)^b \right) \]
\[ = \arg\max_p a \log p + b \log(1 - p) \]

Calculus 101: Set the derivative to zero

\[ P_{best} = \frac{b}{a + b} \]
The “learning” algorithm

Say you have $a$ Work and $b$ Not-Work

$$p_{\text{best}} = \arg\max_p P(D|h)$$

$$= \arg\max_p \log P(D|h)$$

$$= \arg\max_p \log \left( \frac{a+b}{a} p^a (1-p)^b \right)$$

$$= \arg\max_p a \log p + b \log (1-p)$$

Calculus 101: Set the derivative to zero

$$P_{\text{best}} = \frac{b}{a+b}$$

The model we assumed is Bernoulli. You could assume a different model! Next we will consider other models and see how to learn their parameters.
Today’s lecture

• Bayesian Learning

• Maximum a posteriori and maximum likelihood estimation

• Two examples of maximum likelihood estimation
  – Bernoulli trials
  – Normal distribution
Maximum Likelihood estimation

Maximum Likelihood estimation (MLE)

\[ h_{ML} = \arg \max_{h \in H} P(D|h) \]

What we need in order to define learning:
1. A hypothesis space \( H \)
2. A model that says how data \( D \) is generated given \( h \)
Example 2:

Maximum Likelihood and least squares

\[ h_{ML} = \arg \max_{h \in H} P(D|h) \]

Suppose \( H \) consists of real valued functions

Inputs are vectors \( \mathbf{x} \in \mathbb{R}^d \) and the output is a real number \( y \in \mathbb{R} \)
Example 2:  
Maximum Likelihood and least squares

\[ h_{ML} = \arg \max_{h \in H} P(D|h) \]

Suppose H consists of real valued functions
Inputs are vectors \( \mathbf{x} \in \mathbb{R}^d \) and the output is a real number \( y \in \mathbb{R} \)

Suppose the training data is generated as follows:
- An input \( x_i \) is drawn randomly (say uniformly at random)
- The true function \( f \) is applied to get \( f(x_i) \)
- This value is then corrupted by noise \( e_i \)
  - Drawn independently according to an unknown Gaussian with zero mean
Example 2:

Maximum Likelihood and least squares

\[ h_{ML} = \arg \max_h P(D|h) \]

Suppose \( H \) consists of real valued functions.
Inputs are vectors \( x \in \mathbb{R}^d \) and the output is a real number \( y \in \mathbb{R} \).

Suppose the training data is generated as follows:

• An input \( x_i \) is drawn randomly (say uniformly at random).
• The true function \( f \) is applied to get \( f(x_i) \).
• This value is then corrupted by noise \( e_i \):
  – Drawn independently according to an unknown Gaussian with zero mean.

\[ y_i = f(x_i) + e_i \]
Example 2:
Maximum Likelihood and least squares

\[ h_{ML} = \arg \max_{h \in H} P(D|h) \]

Suppose \( H \) consists of real valued functions
Inputs are vectors \( x \in \mathbb{R}^d \) and the output is a real number \( y \in \mathbb{R} \)

Suppose the training data is generated as follows:
- An input \( x_i \) is drawn randomly (say uniformly at random)
- The true function \( f \) is applied to get \( f(x_i) \)
- This value is then corrupted by noise \( e_i \)
  - Drawn independently according to an unknown Gaussian with zero mean

\[ y_i = f(x_i) + e_i \]

Say we have \( m \) training examples \((x_i, y_i)\) generated by this process
Example:
Maximum Likelihood and least squares

Suppose we have a hypothesis $h$. We want to know what is the probability that a particular label $y_i$ was generated by this hypothesis as $h(x_i)$?

The error for this example is $y_i - h(x_i)$.
Suppose we have a hypothesis $h$. We want to know what is the probability that a particular label $y_i$ was generated by this hypothesis as $h(x_i)$?

The error for this example is $y_i - h(x_i)$

We believe that this error is from a Gaussian distribution with mean = 0 and some standard deviation (as we will see, we do not need to know it)

We can compute the probability of observing one data point $(x_i, y_i)$, if it were generated using the function $h$
Maximum Likelihood and least squares

Probability of observing one data point \((x_i, y_i)\), if it were generated using the function \(h\)

\[
p(y_i|h, x_i) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i - h(x_i))^2}{2\sigma^2}}
\]
Example:

Maximum Likelihood and least squares

Probability of observing one data point \((x_i, y_i)\), if it were generated using the function \(h\)

\[
p(y_i|h, x_i) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i - h(x_i))^2}{2\sigma^2}}
\]

Each example in our dataset \(D = \{(x_i, y_i)\}\) is generated \textit{independently} by this process

\[
p(D|h) = \prod_{i=1}^{m} p(y_i, x_i|h) \propto \prod_{i=1}^{m} p(y_i|h, x_i)
\]
Example:

Maximum Likelihood and least squares

\[ h_{ML} = \arg \max_{h \in H} P(D|h) \]

Our goal is to find the most likely hypothesis

\[ h_{ML} = \arg \max_{h \in H} p(D|h) = \arg \max_{h \in H} \prod_{i=1}^{m} p(y_i|h, x_i) \]
Example:

Maximum Likelihood and least squares

$$h_{ML} = \arg \max_{h \in H} P(D|h)$$

Our goal is to find the most likely hypothesis

$$h_{ML} = \arg \max_{h \in H} p(D|h) = \arg \max_{h \in H} \prod_{i=1}^{m} p(y_i|h, x_i)$$

$$= \arg \max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i-h(x_i))^2}{2\sigma^2}}$$
Maximum Likelihood and least squares

$$h_{ML} = \arg \max_{h \in H} P(D|h)$$

Our goal is to find the most likely hypothesis

$$h_{ML} = \arg \max_{h \in H} p(D|h) = \arg \max_{h \in H} \prod_{i=1}^{m} p(y_i|h, x_i)$$

$$= \arg \max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i-h(x_i))^2}{2\sigma^2}}$$

Example:
How do we maximize this expression? Any ideas?
Example:

**Maximum Likelihood and least squares**

\[ h_{ML} = \arg \max_{h \in H} P(D|h) \]

Our goal is to find the most likely hypothesis

\[
h_{ML} = \arg \max_{h \in H} p(D|h) = \arg \max_{h \in H} \prod_{i=1}^{m} p(y_i|h, x_i) \\
= \arg \max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i - h(x_i))^2}{2\sigma^2}}
\]

How do we maximize this expression? Any ideas?

**Answer:** Take the logarithm to simplify
Example:
Maximum Likelihood and least squares

\[ h_{ML} = \arg \max_{h \in H} P(D|h) \]

Our goal is to find the most likely hypothesis

\[
\begin{align*}
  h_{ML} &= \arg \max_{h \in H} p(D|h) = \arg \max_{h \in H} \prod_{i=1}^{m} p(y_i|h, x_i) \\
  &= \arg \max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i - h(x_i))^2}{2\sigma^2}} \\
  &= \arg \max_{h \in H} \sum_{i=1}^{m} \log \frac{1}{\sigma \sqrt{2\pi}} - \frac{(y_i - h(x_i))^2}{2\sigma^2}
\end{align*}
\]
Example:

**Maximum Likelihood and least squares**

\[ h_{ML} = \arg \max_{h \in H} P(D|h) \]

Our goal is to find the most likely hypothesis

\[
h_{ML} = \arg \max_{h \in H} p(D|h) = \arg \max_{h \in H} \prod_{i=1}^{m} p(y_i|h, x_i) \]

\[
= \arg \max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i - h(x_i))^2}{2\sigma^2}} \]

\[
= \arg \max_{h \in H} \sum_{i=1}^{m} \log \frac{1}{\sigma \sqrt{2\pi}} - \frac{(y_i - h(x_i))^2}{2\sigma^2} \]

\[
= \arg \max_{h \in H} - \sum_{i=1}^{m} \frac{(y_i - h(x_i))^2}{2\sigma^2} \]
Example:

Maximum Likelihood and least squares

\[ h_{ML} = \arg \max_{h \in H} P(D|h) \]

Our goal is to find the most likely hypothesis

\[ h_{ML} = \arg \max_{h \in H} p(D|h) = \arg \max_{h \in H} \prod_{i=1}^{m} p(y_i|h, x_i) \]

\[ = \arg \max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_i-h(x_i))^2}{2\sigma^2}} \]

\[ = \arg \max_{h \in H} \sum_{i=1}^{m} \log \frac{1}{\sigma \sqrt{2\pi}} - \frac{(y_i-h(x_i))^2}{2\sigma^2} \]

\[ = \arg \max_{h \in H} - \sum_{i=1}^{m} \frac{(y_i-h(x_i))^2}{2\sigma^2} \]

\[ = \arg \min_{h \in H} \sum_{i=1}^{m} (y_i-h(x_i))^2 \]

Because we assumed that the standard deviation is a constant.
The most likely hypothesis is

\[ h_{ML} = \arg \min_{h \in H} \sum_{i=1}^{m} (y_i - h(x_i))^2 \]
Example:

Maximum Likelihood and least squares

The most likely hypothesis is

$$h_{ML} = \arg \min_{h \in H} \sum_{i=1}^{m} (y_i - h(x_i))^2$$

If we consider the set of linear functions as our hypothesis space: $h(x_i) = w^T x_i$

$$h_{ML} = \arg \min_{w} \sum_{i=1}^{m} (y_i - w^T x_i)^2$$
Example:

Maximum Likelihood and least squares

The most likely hypothesis is

$$h_{ML} = \arg \min_{h \in H} \sum_{i=1}^{m} (y_i - h(x_i))^2$$

If we consider the set of linear functions as our hypothesis space: $h(x_i) = w^T x_i$

$$h_{ML} = \arg \min_{w} \sum_{i=1}^{m} (y_i - w^T x_i)^2$$

This is the probabilistic explanation for least squares regression
Linear regression: Two perspectives

Loss minimization perspective
We want to minimize the difference between the squared loss error of our prediction

Minimize the total squared loss

Bayesian perspective
We believe that the errors are Normally distributed with zero mean and a fixed variance

Find the linear regressor using the maximum likelihood principle

\[
\arg\min_w \sum_{i=1}^{m} (y_i - w^T x_i)^2
\]
Today’s lecture: Summary

• Bayesian Learning
  – Another way to ask: What is the best hypothesis for a dataset?
  – Two answers to the question: Maximum a posteriori (MAP) and maximum likelihood estimation (MLE)

• We saw two examples of maximum likelihood estimation
  – Binomial distribution, normal distribution
  – You should be able to apply both MAP and MLE to simple problems