Machine Learning

Expectation Maximization
(and Probability Review)

Zach Wood-Doughty and Bryan Pardo, CS349 Fall 2021
Axioms of Probability

- Let there be a space $S$ composed of a countable number of events
  
  $$S \equiv \{e_1, e_2, e_3, \ldots, e_n\}$$

- The probability of each event is between 0 and 1
  
  $$0 \leq P(e_1) \leq 1$$

- The probability of the whole sample space is 1
  
  $$P(S) = 1$$

- **When two events are mutually exclusive, their probabilities are additive**
  
  $$P(e_1 \lor e_2) = P(e_1) + P(e_2)$$
Discrete Random Variables

- $P(\text{Grade})$ is a distribution over possible grades
- Each grade is mutually exclusive
- Probabilities sum to 1
Boolean Random Variable

- Boolean random variable: A random variable that has only two possible outcomes
  e.g.

  \( X = \text{“Tomorrow’s high temperature > 60”} \) has only two possible outcomes

  As a notational convention, \( P(X) \) for a Boolean variable will mean \( P(X=\text{“true”}) \), since it is easy to infer the rest of the distribution.
Vizualizing $P(A)$ for a Boolean variable

$$P(A) = \frac{\text{area of yellow oval}}{\text{area of blue rectangle}}$$

0 $\leq P(A) \leq 1$

If a value is over 1 or under 0, it isn't a probability
Visualizing two Booleans

\[ P(A \lor B) = P(A) + P(B) - P(A \land B) \]
Independence

• variables A and B are said to be \textit{independent} iff...

\[ P(A)P(B) = P(A \land B) \]
Bayes Rule

• Definition of Conditional Probability

\[ P(A \mid B) = \frac{P(A \land B)}{P(B)} \]

• Corollary: The Chain Rule

\[ P(A \mid B)P(B) = P(A \land B) \]

• Bayes Rule

(Thomas Bayes, 1763)

\[ P(B \mid A) = \frac{P(A \land B)}{P(A)} = \frac{P(A \mid B)P(B)}{P(A)} \]
Conditional Probability

The conditional probability of A given B is represented by the following formula:

\[ P(A \mid B) = \frac{P(A \cap B)}{P(B)} \]

Overlap implies NOT independent

Can we do the following?

\[ P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} \]

Only if A and B are independent

Zach Wood-Doughty and Bryan Pardo, CS349 Fall 2021
The Joint Distribution

- Truth table lists all combinations of variable assignments
- Assign a probability to each row
- Probabilities sum to 1

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Using The Joint Distribution

- Find $P(A)$
- Sum the probabilities of all rows where $A=1$

\[
P(A) = 0.05 + 0.2 + 0.25 + 0.05 = 0.55
\]
Using The Joint Distribution

- Find \( P(A|B) \)

\[
p(A \mid B) = \frac{p(A, B)}{p(B)}
\]

\[
p(B = b) = \sum_{a \in \{0, 1\}} p(A = a, B = b)
\]

\[
= (0.25 + 0.05) \div (0.25 + 0.05 + 0.1 + 0.05)
\]

\[
= 0.3 \div 0.45
\]

\[
= 0.667
\]
### Using The Joint Distribution

Are A and B Independent?

- \( P(A, B) = 0.25 + 0.05 \)
- \( P(A) = 0.3 + 0.2 + 0.05 \)
- \( P(B) = 0.3 + 0.1 + 0.05 \)

\[
P(A) \times P(B) = 0.55 \times 0.45 = 0.2475
\]

\( P(A, B) = 0.3 \neq 0.248 \)

A and B NOT independent

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Why not use the Joint Distribution?

- Given $m$ boolean variables, we need to estimate $2^m$ values.

- 20 yes-no questions = a million values

- How do we get around this combinatorial explosion?
  - Assume independence of variables!
...back to independence

- The probability I eat pie today is independent of the probability of a blizzard in Japan.
- This is DOMAIN knowledge, typically supplied by the problem designer.
- Independence implies:

\[ A \perp B \Rightarrow p(A \mid B) = p(A) \]
\[ A \perp B \mid C \Rightarrow p(A, B \mid C) = p(A \mid C)p(B \mid C) \]
Let’s show that

assuming independence...

\[ P(A \land B) = P(A)P(B) \]

plus the chain rule...

\[ P(A \land B) = P(A \mid B)P(B) \]

imply...

\[ P(A)P(B) = P(A \mid B)P(B) \]

which means...

\[ P(A \mid B) = P(A) \]
Some Definitions

• Prior probability of $h$, $P(h)$:
  – background knowledge on probability that $h$ is a correct hypothesis (before having observed the data)

• Conditional Probability of $D$, $P(D|h)$:
  – the probability of observing data $D$ given that hypothesis $h$ holds

• Posterior probability of $h$, $P(h|D)$:
  – the probability of, given the observed training data $D$
  – this is what we want!
Maximum A Posteriori (MAP)

- **Goal:** To find the most probable hypothesis $h$ from a set of candidate hypotheses $H$ given the observed data $D$.

- **MAP Hypothesis, $h_{\text{MAP}}$**

\[
\begin{align*}
  h_{\text{map}} &= \arg\max_{h \in H} (P(h \mid D)) \\
  &= \arg\max_{h \in H} \left( \frac{P(D \mid h)P(h)}{P(D)} \right) \\
  &= \arg\max_{h \in H} (P(D \mid h)P(h))
\end{align*}
\]
Maximum Likelihood (ML)

- **ML hypothesis** is a special case of the MAP hypothesis where all hypotheses are, to begin with, equally likely.

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

Assume...

\[
P(h) = \frac{1}{|H|} \quad \forall h \in H
\]

Then...

\[
h_{\text{ml}} = \arg \max_{h \in H} (P(D | h))
\]
MAP vs Maximum Likelihood

\[ P(\text{cancer}) = 0.01 \]
\[ P(\text{positive test} \mid \text{cancer}) = 0.97 \]
\[ P(\text{positive test} \mid \text{no cancer}) = 0.02 \]

What is \( p(\text{cancer} \mid \text{positive test}) \)?
Base Rate Fallacy

Total Population = 100 people; 83% vaccination rate

Vaccinated

Unvaccinated

50% of infections were among vaccinated

yourlocalepidemiologist.substack.com

xkcd.com/2476/

REMEMBER, RIGHT-HANDED PEOPLE COMMIT 90% OF ALL BASE RATE ERRORS.
Observed \((x, y)\) is the combination of a point on the regression line plus noise.

\[
\mathbf{w}_{\text{MAP}} = \arg \max_w p(\mathbf{w} | \mathbf{X}, \mathbf{y})
\]

\[
= \arg \max_w p(\mathbf{X}, \mathbf{y} | \mathbf{w}) p(\mathbf{w})
\]
Linear Regression, Again

\[
p(\langle x_i, y_i \rangle; w) = \mathcal{N}(y_i; \mu = w^\top x_i, \sigma = \sigma)
\]

\[
\log p(X, y \mid w, \sigma) = \log \prod_{i=1}^{N} \mathcal{N}(y_i; \mu = w^\top x_i, \sigma = \sigma)
\]

\[
= -\frac{N}{2} \log(2\pi \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - w^\top x_i)^2
\]

\[
w^* = \arg \max_w \log p(w \mid X, y, \sigma)
\]

\[
= \arg \max_w (\log p(X, y \mid w, \sigma) + \log p(w))
\]

\[
= \arg \max_w \left(-\frac{N}{2} \log(2\pi \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - w^\top x_i)^2 + \log p(w)\right)
\]
Linear Regression, Again

$$\log p(\mathbf{X}, \mathbf{y} \mid \mathbf{w}, \sigma) = \log \prod_{i=1}^{N} \mathcal{N}(y_i; \mu = \mathbf{w}^\top \mathbf{x}_i, \sigma = \sigma)$$

$$= -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2$$

$$0 = \frac{d}{d\mathbf{w}} \left( -\frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \right)$$

$$= \left( \sum_{i=1}^{N} y_i \mathbf{x}_i^\top \right) - \mathbf{w}^\top \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top$$

$$= \mathbf{X}^\top \mathbf{y} - \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}$$

$$= \ldots = \mathbf{w} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$
For linear regression, minimizing loss and maximizing likelihood are equivalent!

\[
L_s(X, Y; \theta) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - h_\theta(x_i))^2 \]

\[-\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - w^\top x_i)^2 \]

But what about that \( p(w) \) term?

\[
\arg \max_w \left( -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - w^\top x_i)^2 + \log p(w) \right) \]
What is $p(w)$ for linear regression?

$$p(w) = \mathcal{N}(0, \lambda^{-1})$$

$$w^* = \arg\max_w \log p(w \mid X, y, \sigma)$$

$$= \arg\max_w (\log p(X, y, \mid w, \sigma) + \log p(w))$$

$$= \arg\max_w \left( -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - w^\top x_i)^2 + \log p(w) \right)$$

$$\Rightarrow \arg\max_w \left( \ldots - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - w^\top x_i)^2 - \frac{1}{2} w^2 \lambda^2 \right)$$

$$L_R(X, Y; \theta) = L(X, Y; \theta) + \lambda R(\theta) \quad R_2(\theta) = \frac{1}{2} \sum_{i=1}^{d} |\theta_i|^2$$

Zach Wood-Doughty and Bryan Pardo, CS349 Fall 2021
Latent Variable Models

\[
\max_w p(Y \mid X; w) = \prod_{i=1}^{n} p(y_i \mid x_i; w)
\]

\[
\max_w p(X; \Theta) = \prod_{i=1}^{n} p(x_i; \Theta)
\]

\[
\max_w p(X; \Theta) = \prod_{i=1}^{n} \sum_{k} p(x_i, z_k; \Theta)
\]
Expectation Maximization

Given joint distribution \( p(X, Z \mid \Theta) \), with \( X \) observed and \( Z \) latent, and parameters \( \Theta \), we want to find a \( \Theta \) that maximizes \( p(X \mid \Theta) \).

First: initialize \( \Theta^0 \). Then, repeat until converged:

1. Estimate \( p(Z \mid X, \theta^t) \)
2. Set \( \theta^{t+1} = \arg\max_{\hat{\theta}} p(Z \mid X, \theta^t) \log p(X, Z \mid \hat{\theta}) \)
EM for Gaussian Mixture Model

(Log) Likelihood of GMM:

\[
p(X, Z | \mu, \Sigma, \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_{n}^{z_{nk}} \mathcal{N}(x_n | \mu_k, \Sigma_k)^{z_{nk}}
\]

\[
\ln p(X, Z | \mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(x_n | \mu_k, \Sigma_k)\}
\]

1. Estimate \( p(Z | X, \theta^t) \)

2. Set \( \theta^{t+1} = \arg \max_{\hat{\theta}} p(Z | X, \theta^t) \log p(X, Z | \hat{\theta}) \)
Gaussian Mixture Model

1. Estimate $\mathbb{P}(Z \mid X, \theta^t)$

2. Set $\theta^{t+1} = \arg \max_{\hat{\theta}} \mathbb{P}(Z \mid X, \theta^t) \log \mathbb{P}(X, Z \mid \hat{\theta})$

Cluster Responsibilities

$$\gamma(z_{n,k}) = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$$

Cluster means, variances, and weight coefficients

$$N_k = \sum_{n=1}^{N} \gamma(z_{n,k})$$

$$\pi_k = \frac{N_k}{N}$$

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{n,k})x_n$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{n,k})(x_n - \mu_k)(x_n - \mu_k)^T$$
In the log likelihood function, or alternatively in the parameters, falls below some threshold, the algorithm stops. The change in the log likelihood, either in the function or in the parameters, falls below some threshold. The algorithm is deemed to have converged when the change is sufficiently small.

The Expectation Maximization (EM) algorithm is an iterative optimization algorithm used to find maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. It alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step.

In the E step, we use the current values for the parameters to evaluate the posterior distribution. In the M step, we first evaluate the new means using the results from the E step, then use these new values to find the new covariances, and finally compute the new weights using the results from the E and M steps. This process is guaranteed to increase the log likelihood function at each iteration, and the algorithm is deemed to have converged when the change in the log likelihood falls below a threshold.

Figure 9.1 shows the data points in green, together with the initial configuration of the mixture model. The one standard-deviation contours for the two Gaussians are shown in blue. Figure 9.8 shows the rescaled Old Faithful data set, with a mixture of two Gaussians used, with centres initialized using the same values as for the previous figure. Here a mixture of two Gaussians is used, with centres initialized using the same values as for the previous figure. The one standard-deviation contours for the two Gaussians are shown in red.

The algorithm in Figure 9.1 is shown in green. In practice, the algorithm is deemed to have converged when the change in the log likelihood falls below some threshold.

For the rescaled Old Faithful data set in Figure 9.8, a mixture of two Gaussians is used, with centres initialized using the same values as for the previous figure. Here a mixture of two Gaussians is used, with centres initialized using the same values as for the previous figure. The one standard-deviation contours for the two Gaussians are shown in red.

The algorithm in Figure 9.1 is shown in green. In practice, the algorithm is deemed to have converged when the change in the log likelihood falls below some threshold.
Semi-supervised Learning
Recall: Supervised Learning Tasks

There is a set of possible examples \( X = \{x_1, \ldots, x_n\} \)

Each example is a vector of \( d \) real valued attributes

\[
x_i = \langle x_{i,1}, \ldots, x_{i,d} \rangle
\]

A target function maps \( X \) onto some real or categorical value \( Y \)

\[
f : X \rightarrow Y
\]

The DATA is a set of tuples \(<\text{example}, \text{response value}>\)

\[
\{<x_1, y_1>, \ldots, <x_n, y_n>\}
\]

Find a hypothesis \( h \) such that...

\[
\forall x, h(x) \approx f(x)
\]
Unsupervised Learning Tasks

There is a set of possible examples

\[ X = \{x_1, \ldots, x_n\} \]

Each example is a **vector** of \( d \) **real valued attributes**

\[ x_i = \langle x_{i,1}, \ldots, x_{i,d} \rangle \]

Assume some latent variable(s) \( z \) that correspond to the observed data

\[ \{\langle x_1, z_1 \rangle, \ldots, \langle x_n, z_n \rangle\} \]

Learn a joint distribution of \( p(X, Z) \)
A necessary condition of semi-supervised learning is that the underlying marginal data distribution never seen the phrase "particle accelerator" in the labelled data.

Towards classifying these documents as revolving around physics as well, despite having "neutron" to the phrase "particle accelerator". For instance, the word "neutron" would often occur in a document that also contains the word "quark". Furthermore, the word "quark" may only appear in documents concerning physics but never in documents concerning particle accelerators that does not contain the word "neutron", the classifier is unable to recognize it as a document concerning physics. This is where semi-supervised learning comes in. If we consider the unlabelled data, there might be documents that connect the word "quark" to the phrase "particle accelerator". For instance, the word "quark" would often occur in a document that also contains the word "neutron" to the phrase "particle accelerator".

Supervised Learning

Information about the posterior distribution over the input space contains information about the posterior distribution. If, on the other hand, this condition is not met, and contains no information about the posterior distribution, it is inherently impossible to improve the accuracy of predictions.

Abasic example of binary classification with unlabelled data. The unlabelled data points provide some further intuition towards the use of unlabelled data for classification.

Semi-Supervised Learning

Animation on Course Website!

Acc = 65.0
Semi-Supervised Learning

Animation on Course Website!

Acc = 65.0
Illustration of the decomposition given by (9.70), which holds for any choice of distribution $q(Z)$. Because the Kullback-Leibler divergence satisfies $\text{KL}(q||p) \geq 0$, we see that the quantity $\mathcal{L}(q, \theta)$ is a lower bound on the log likelihood function $\ln p(X|\theta)$.

\[ \ln p(X|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q||p) \]

\[ \mathcal{L}(q, \theta) = \sum_Z q(Z) \ln \left\{ \frac{p(X, Z|\theta)}{q(Z)} \right\} \]

\[ \text{KL}(q||p) = -\sum_Z q(Z) \ln \left\{ \frac{p(Z|X, \theta)}{q(Z)} \right\} \]
**EM: Pictorial View**

Illustration of the E step of the EM algorithm. The $q$ distribution is set equal to the posterior distribution for the current parameter values $\theta^{\text{old}}$, causing the lower bound to move up to the same value as the log likelihood function, with the KL divergence vanishing.

\[
\text{KL}(q||p) = 0
\]

\[
\mathcal{L}(q, \theta^{\text{old}}) = \ln p(X|\theta^{\text{old}})
\]

\[
\mathcal{L}(q, \theta) = \sum_Z p(Z|X, \theta^{\text{old}}) \ln p(X, Z|\theta) - \sum_Z p(Z|X, \theta^{\text{old}}) \ln p(Z|X, \theta^{\text{old}})
\]

\[
= Q(\theta, \theta^{\text{old}}) + \text{const}
\]

(9.74)
Illustration of the M step of the EM algorithm. The distribution \( q(Z) \) is held fixed and the lower bound \( \mathcal{L}(q, \theta) \) is maximized with respect to the parameter vector \( \theta \) to give a revised value \( \theta^{\text{new}} \). Because the KL divergence is nonnegative, this causes the log likelihood \( \ln p(X|\theta) \) to increase by at least as much as the lower bound does.
The EM algorithm involves alternating between the E (Estimation) step, where the posterior distribution over latent variables is evaluated, giving a lower bound \( L(q, \theta) \) whose value equals the log likelihood at \( \theta^{\text{old}} \), as shown by the blue curve. Note that the bound makes a tangential contact with the log likelihood at \( \theta^{\text{old}} \), so that both curves have the same gradient. This bound is a convex function having a unique maximum (for mixture components from the exponential family). In the M (Maximization) step, the bound is maximized giving the value \( \theta^{\text{new}} \), which gives a larger value of log likelihood than \( \theta^{\text{old}} \). The subsequent E step then constructs a bound that is tangential at \( \theta^{\text{new}} \) as shown by the green curve.

For the particular case of an independent, identically distributed data set, \( X \) will comprise \( N \) data points \( \{x_n\} \) while \( Z \) will comprise \( N \) corresponding latent variables \( \{z_n\} \), where \( n = 1, \ldots, N \). From the independence assumption, we have

\[
p(X, Z) = \prod_{n=1}^{N} p(x_n, z_n)
\]

and, by marginalizing over the \( \{z_n\} \) we have

\[
p(X) = \prod_{n=1}^{N} p(x_n)
\]

Using the sum and product rules, we see that the posterior probability that is evaluated in the E step takes the form

\[
p(Z | X, \theta) = \frac{p(X, Z | \theta)}{\sum_{Z} p(X, Z | \theta)} = \frac{1}{N} \prod_{n=1}^{N} p(x_n, z_n | \theta)
\]

and so the posterior distribution also factorizes with respect to \( n \). In the case of the Gaussian mixture model this simply says that the responsibility that each of the mixture components takes for a particular data point \( x_n \) depends only on the value of \( x_n \) and on the parameters \( \theta \) of the mixture components, not on the values of the other data points.

We have seen that both the E and the M steps of the EM algorithm are increasing the value of a well-defined bound on the log likelihood function and that the

\[
\log p(X | \theta) = L(q, \theta) + KL(q || p)
\]

Increases Can only increase

\[
\log p(X | \theta) \geq \log p(X | \theta^{\text{old}})
\]