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

$$S \equiv \{e_1, e_2, e_3, ...e_n\}$$

- The probability of each event is between 0 and 1

$$0 \leq P(e_1) \leq 1$$

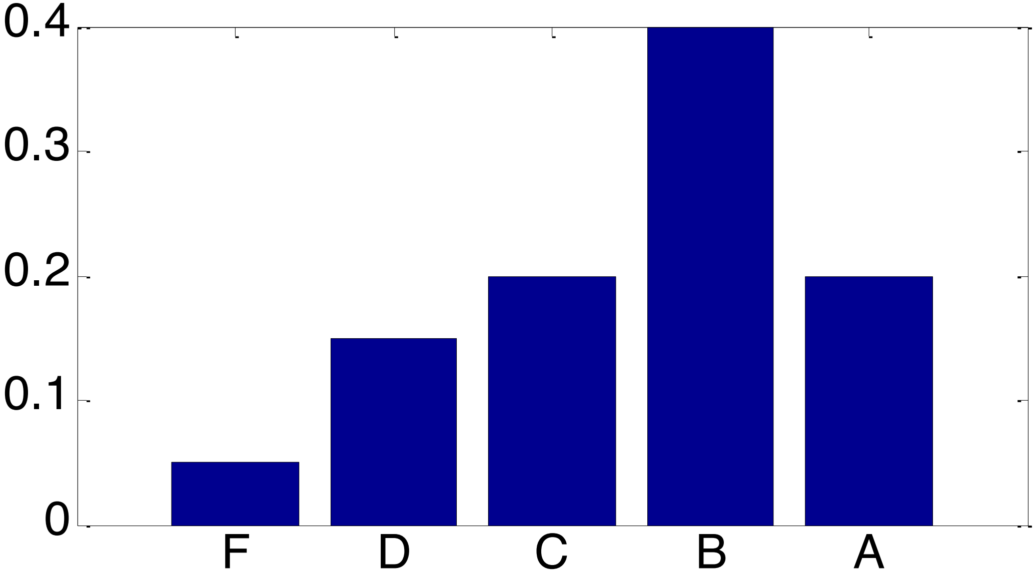
- The probability of the whole sample space is 1

$$P(S) = 1$$

- **When two events are mutually exclusive**, their probabilities are additive

$$P(e_1 \vee e_2) = P(e_1) + P(e_2)$$

Discrete Random Variables



Grade	Probability
A	0.2
B	0.4
C	0.2
D	0.15
F	0.05

- P(Grade) is a distribution over possible grades
- Each grade is mutually exclusive
- Probabilities sum to 1

Boolean Random Variable

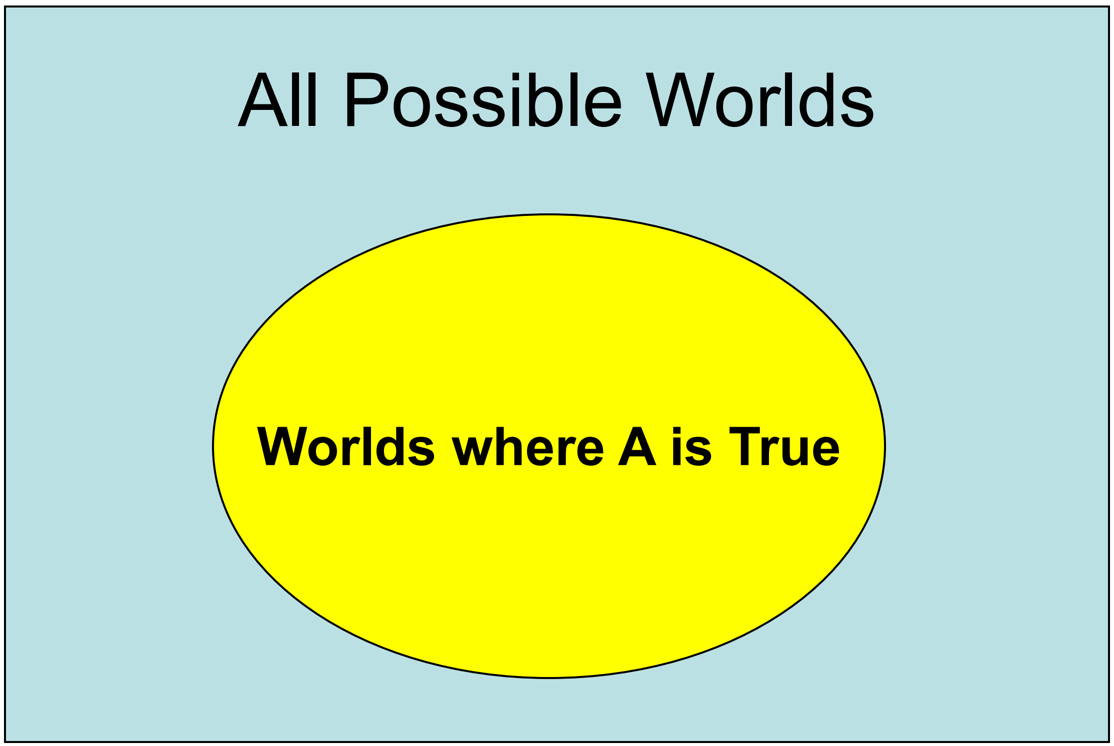
- Boolean random variable: A random variable that has only two possible outcomes

  e.g.

  X = “Tomorrow’s high temperature > 60” has only two possible outcomes

  As a notational convention, **P(X)** for a Boolean variable will mean **P(X=“true”)**, since it is easy to infer the rest of the distribution.

Vizualizing P(A) for a Boolean variable

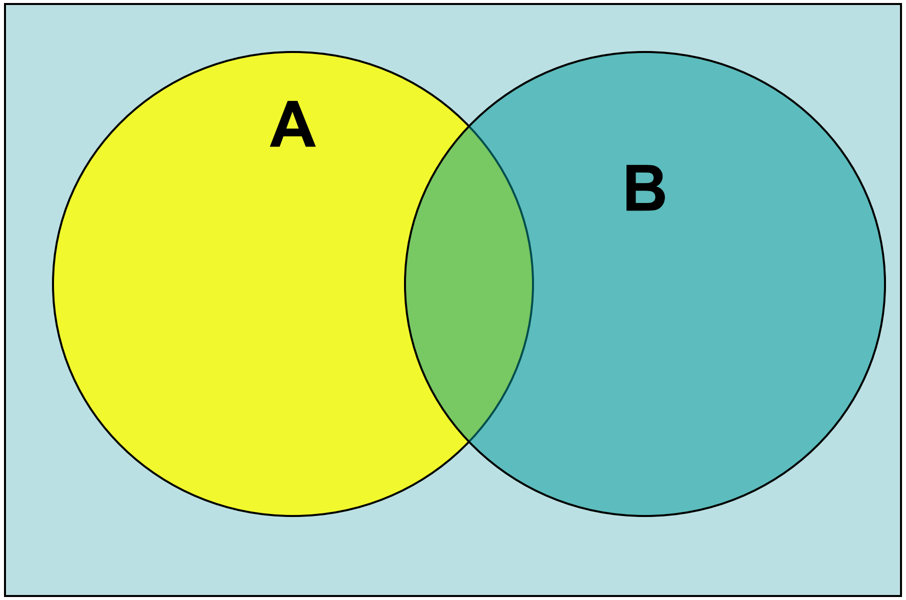


$0 \le P(A) \le 1$

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

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

Visualizing two Booleans



$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

Independence

- variables A and B are said to be *independent* iff…

$$P(A)P(B) = P(A \wedge B)$$

Bayes Rule

- Definition of Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

- Corollary:
  The Chain Rule

$$P(A \mid B)P(B) = P(A \wedge B)$$

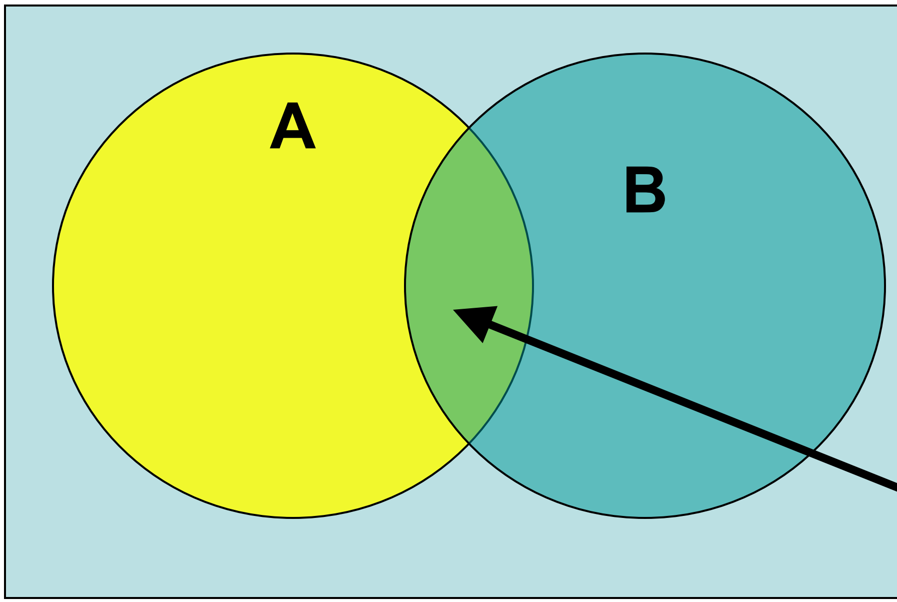
- Bayes Rule

  (Thomas Bayes, 1763)

$$P(B \mid A) = \frac{P(A \wedge B)}{P(A)}$$

$$= \frac{P(A \mid B)P(B)}{P(A)}$$

Conditional Probability



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

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

Overlap implies NOT independent

Can we do the following?

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} = \frac{P(A)P(B)}{P(B)}$$

Only if A and B are ***independent***

The Joint Distribution

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

A	B	C	Prob
0	0	0	0.1
0	0	1	0.2
0	1	0	0.1
0	1	1	0.05
1	0	0	0.05
1	0	1	0.2
1	1	0	0.25
1	1	1	0.05

Using The Joint Distribution

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

$$P(A) = 0.05 + 0.2 + 0.25 + 0.05 = 0.55$$

A	B	C	Prob
0	0	0	0.1
0	0	1	0.2
0	1	0	0.1
0	1	1	0.05
1	0	0	0.05
1	0	1	0.2
1	1	0	0.25
1	1	1	0.05

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)

÷ (0.25+0.05 + 0.1+0.05)

= 0.3 ÷ 0.45

= 0.667

A	B	C	Prob
0	0	0	0.1
0	0	1	0.2
0	1	0	0.1
0	1	1	0.05
1	0	0	0.05
1	0	1	0.2
1	1	0	0.25
1	1	1	0.05

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)×P(B) = 0.55 × 0.45

P(A, B) = 0.3 ≠ 0.248

A and B NOT independent

A	B	C	Prob
0	0	0	0.1
0	0	1	0.2
0	1	0	0.1
0	1	1	0.05
1	0	0	0.05
1	0	1	0.2
1	1	0	0.25
1	1	1	0.05

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...

$$P(A \wedge B) = P(A)P(B)$$

plus the chain rule...

$$P(A \wedge B) = P(A \mid B)P(B)$$

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)

- **<u>Goal:</u>** To find the most probable hypothesis *h* from a set of candidate hypotheses *H* given the observed data *D*.
- ***MAP Hypothesis, h_{MAP}***

$$h_{map} = \underset{h \in H}{\arg\max}(P(h \mid D))$$

$$= \underset{h \in H}{\arg\max}\left(\frac{P(D \mid h)P(h)}{P(D)}\right)$$

$$= \underset{h \in H}{\arg\max}(P(D \mid h)P(h))$$

Maximum Likelihood (ML)

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

$$h_{map} = \arg\max_{h \in H}(P(D \mid h)P(h))$$

Assume...

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

Then...

$$h_{ml} = \arg\max_{h \in H}(P(D \mid h))$$

MAP vs Maximum Likelihood

P(cancer) = 0.01
P(positive test | cancer) = 0.97
P(positive test | no cancer) = 0.02

What is p(cancer | positive test)?

☝️: 5% ✌️: 25% 🤘: 55%
✊: 75% 🖐: 95%

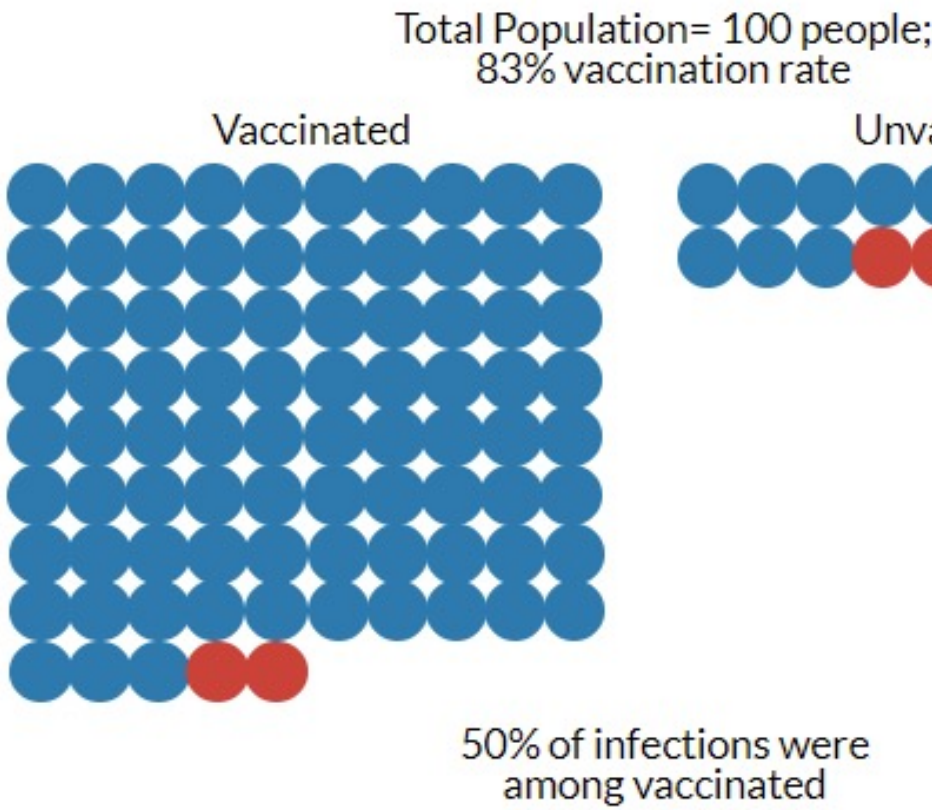
$$p(C \mid T) = \frac{p(T \mid C)p(C)}{p(T)}$$

$$p(T) = p(T \mid C)p(C) + p(T \mid \neg C)p(\neg C)$$
$$= 0.97 \cdot 0.01 + 0.02 \cdot 0.99$$

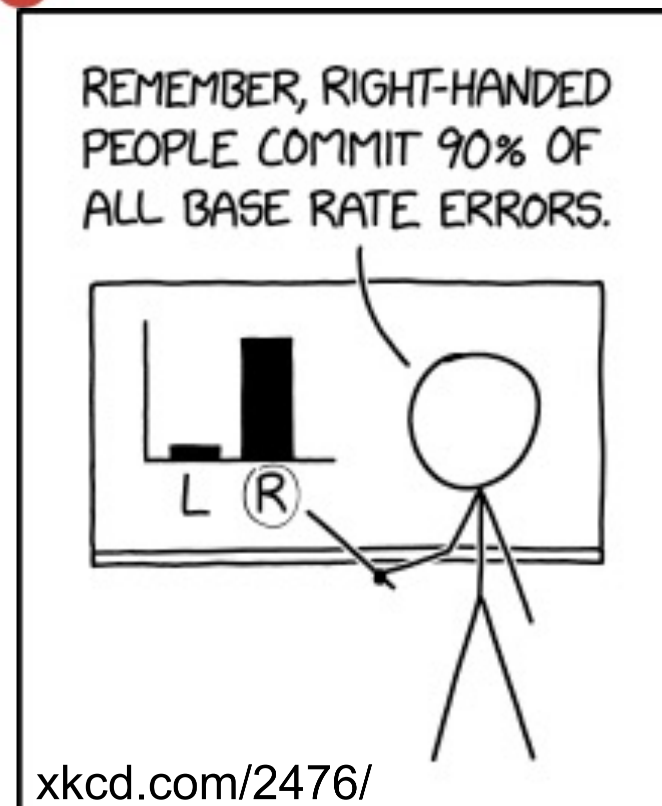
$$p(C \mid T) = \frac{0.97 \cdot 0.01}{0.97 \cdot 0.01 + 0.02 \cdot 0.99}$$

$$= \frac{0.0097}{0.0097 + 0.0198} \approx 0.33\%$$

Base Rate Fallacy

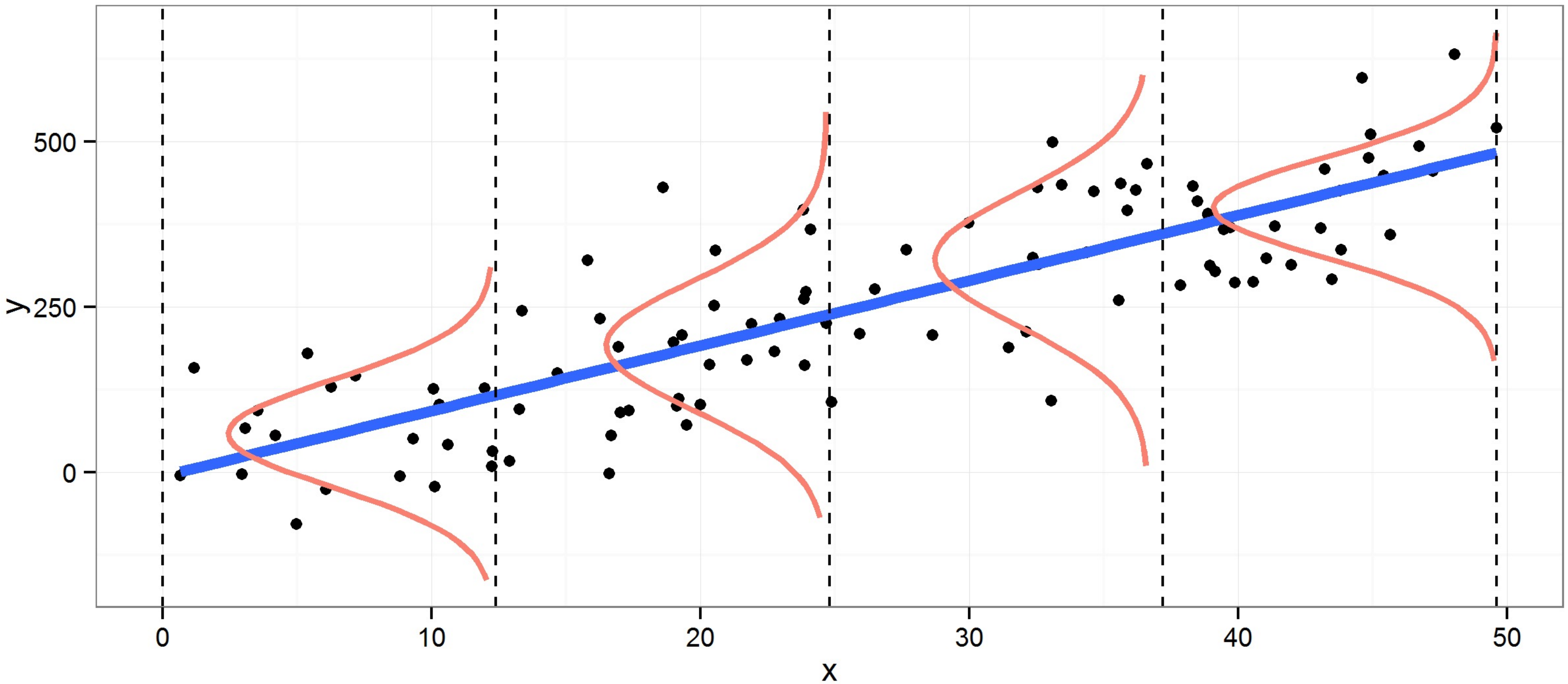


yourlocalepidemiologist.substack.com



xkcd.com/2476/

Linear Regression, Again



Observed (x, y) is the combination of a point on the regression line plus noise.

$$\mathbf{w}_{\text{MAP}} = \arg\max_{w} p(\mathbf{w} \mid \mathbf{X}, \mathbf{y})$$
$$= \arg\max_{w} p(\mathbf{X}, \mathbf{y} \mid \mathbf{w}) p(\mathbf{w})$$

What is p(X, y | w)? p(W)?

https://stackoverflow.com/questions/31794876/ggplot2-how-to-curve-small-gaussian-densities-on-a-regression-line

Linear Regression, Again

$$p(\langle x_i, y_i \rangle; \mathbf{w}) = \mathcal{N}(y_i; \mu = \mathbf{w}^\top \mathbf{x}_i, \sigma = \sigma)$$

$$\begin{aligned}\log p(\mathbf{X}, \mathbf{y} \mid \mathbf{w}, \sigma) &= \log \prod_{i=1}^{N} \mathcal{N}(y_i; \mu = \mathbf{w}^\top \mathbf{x}_i, \sigma = \sigma) \\ &= -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2\end{aligned}$$

$$\begin{aligned}\mathbf{w}^* &= \arg\max_w \log p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}, \sigma) \\ &= \arg\max_w (\log p(\mathbf{X}, \mathbf{y}, \mid \mathbf{w}, \sigma) + \log p(\mathbf{w})) \\ &= \arg\max_w \left(-\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \log p(\mathbf{w}) \right)\end{aligned}$$

Linear Regression, Again

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

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

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

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

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

$$= \ldots = \mathbf{w} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Linear Regression, Again

For linear regression,
minimizing loss and maximizing likelihood are equivalent!

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

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

But what about that p(w) term?

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

What is p(w) for linear regression?

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

$$\begin{aligned}
\mathbf{w}^* &= \arg\max_w \log p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}, \sigma) \\
&= \arg\max_w (\log p(\mathbf{X}, \mathbf{y}, \mid \mathbf{w}, \sigma) + \log p(\mathbf{w})) \\
&= \arg\max_w \left(-\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \log p(\mathbf{w}) \right) \\
&\Rightarrow \arg\max_w \left(\ldots - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 - \frac{1}{2} \mathbf{w}^2 \lambda^2) \right)
\end{aligned}$$

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

Latent Variable Models

$$\max_w p(Y|X;w) = \prod_{i=1}^{n} p(y_i|x_i;w)$$

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

$$\max_w p(X;\Theta) = \prod_{i=1}^{n} \sum_k p(x_i, z_k;\Theta)$$

Expectation Maximization

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

First: initialize Θ^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:

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left\{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}$$

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

Gaussian Mixture Model

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

Cluster Responsibilities

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

Cluster means, variances, and weight coefficients

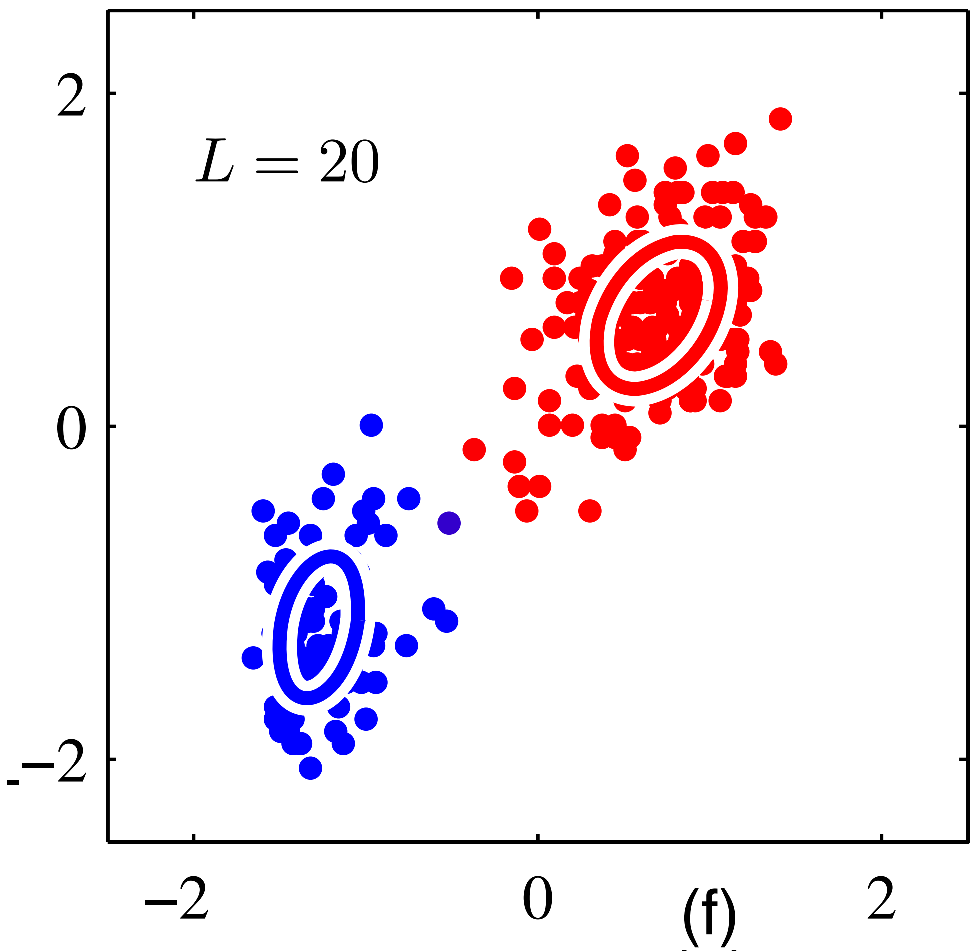
$$N_k = \sum_{n=1}^{N} \gamma(z_{n,k})$$

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

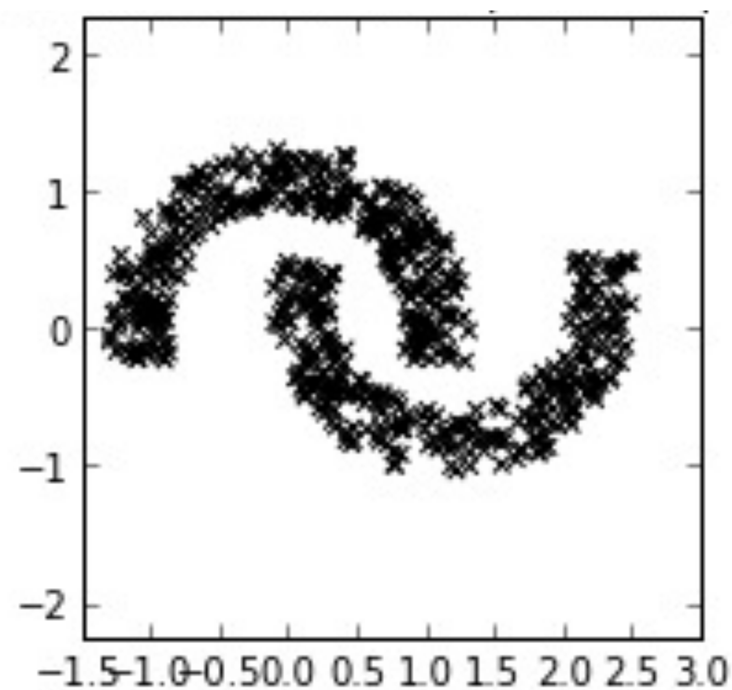
$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{n,k}) \mathbf{x}_n$$

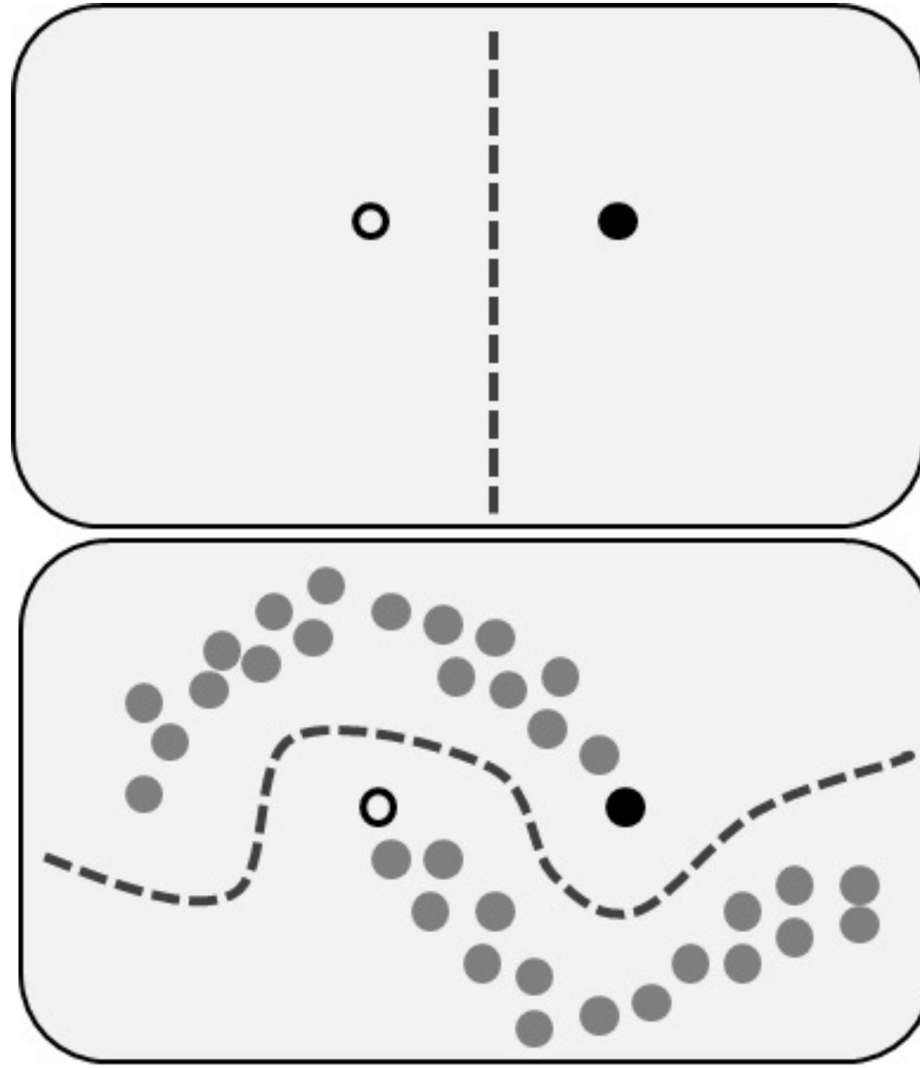
$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{n,k}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^\top$$

Expectation Maximization



Semi-supervised Learning





Recall: Supervised Learning Tasks

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

Each example is a **vector** of d **real valued attributes**

$$\mathbf{x_i} = \langle x_{i,1}, \ldots x_{i,d} \rangle$$

A target function maps *X* onto some **real or categorical value** *Y*

$$f : X \rightarrow Y$$

The DATA is a set of tuples <example, response value>

$$\{< \mathbf{x}_1, y_1 >, \ldots < \mathbf{x_n}, y_n >\}$$

Find a hypothesis ***h*** such that...

$$\forall \mathbf{x}, h(\mathbf{x}) \approx f(\mathbf{x})$$

Unsupervised Learning Tasks

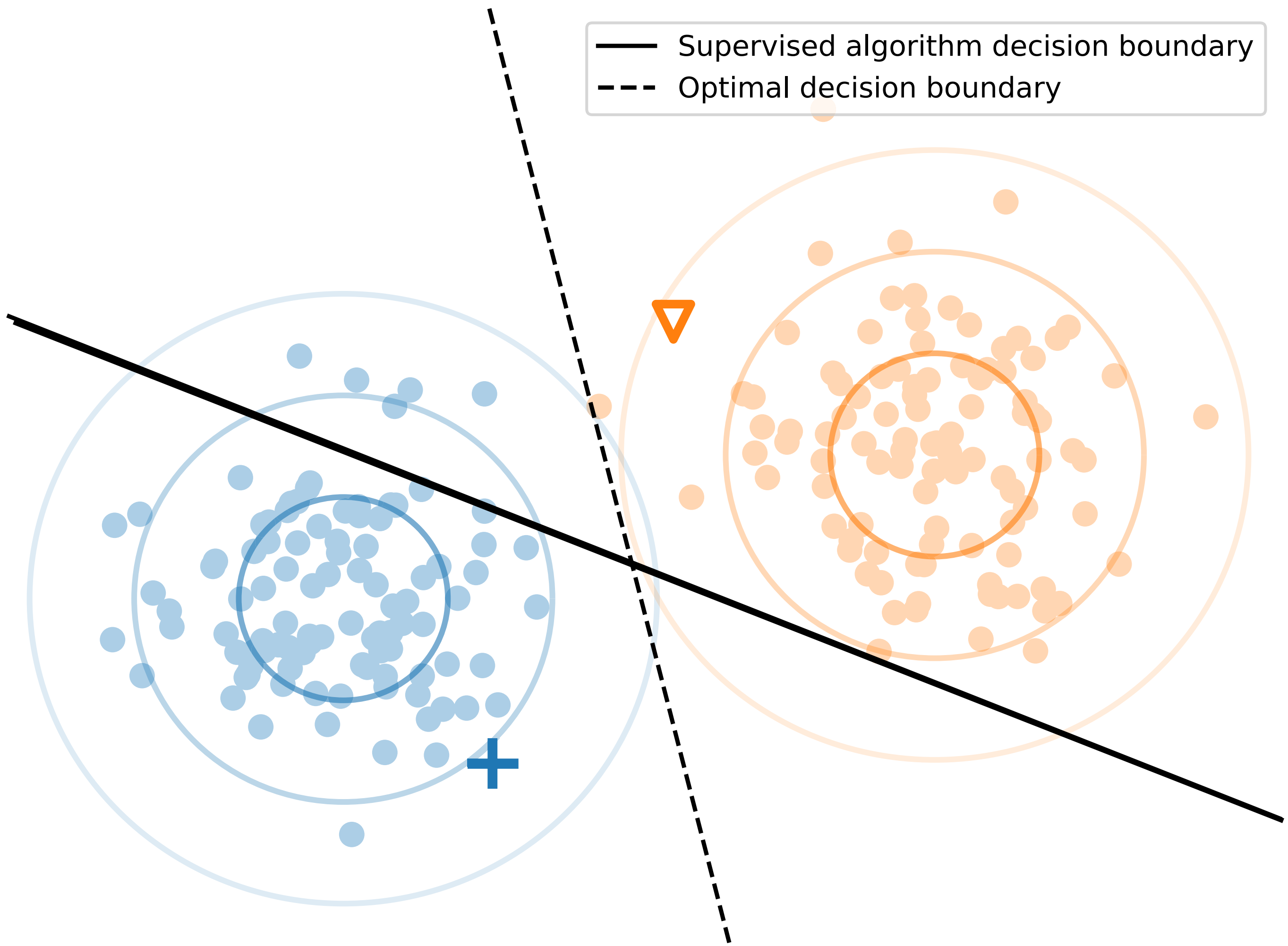
There is a set of possible examples $X = \{\mathbf{x_1}, \ldots \mathbf{x_n}\}$

Each example is a **vector** of d **real valued attributes** $\mathbf{x_i} = \langle x_{i,1}, \ldots x_{i,d} \rangle$

Assume some latent variable(s) z that correspond to the observed data $\{\langle \mathbf{x_1}, z_1 \rangle, \ldots \langle \mathbf{x_n}, z_n \rangle\}$

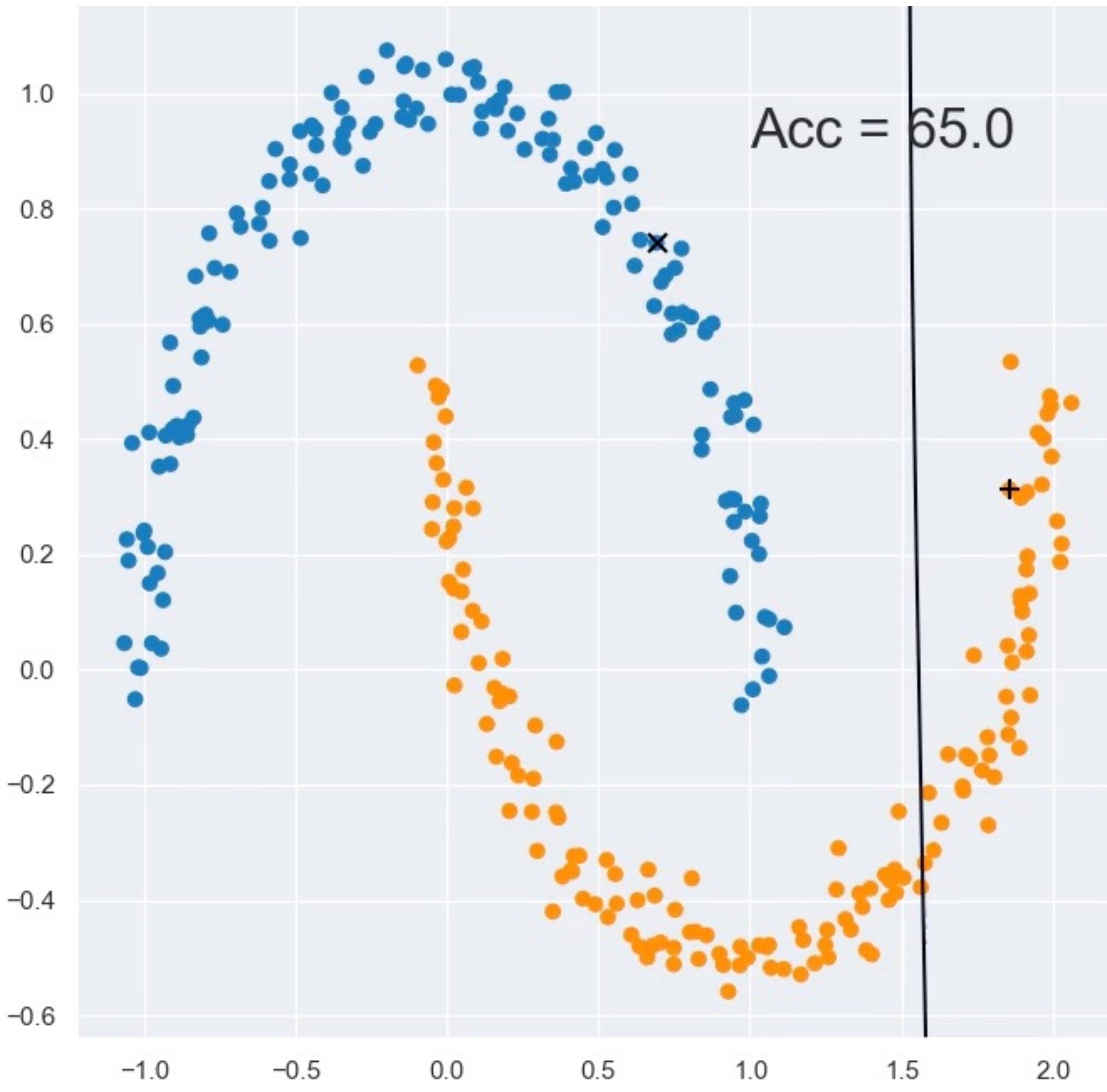
Learn a joint distribution of p(X, Z)

Semi-Supervised Learning

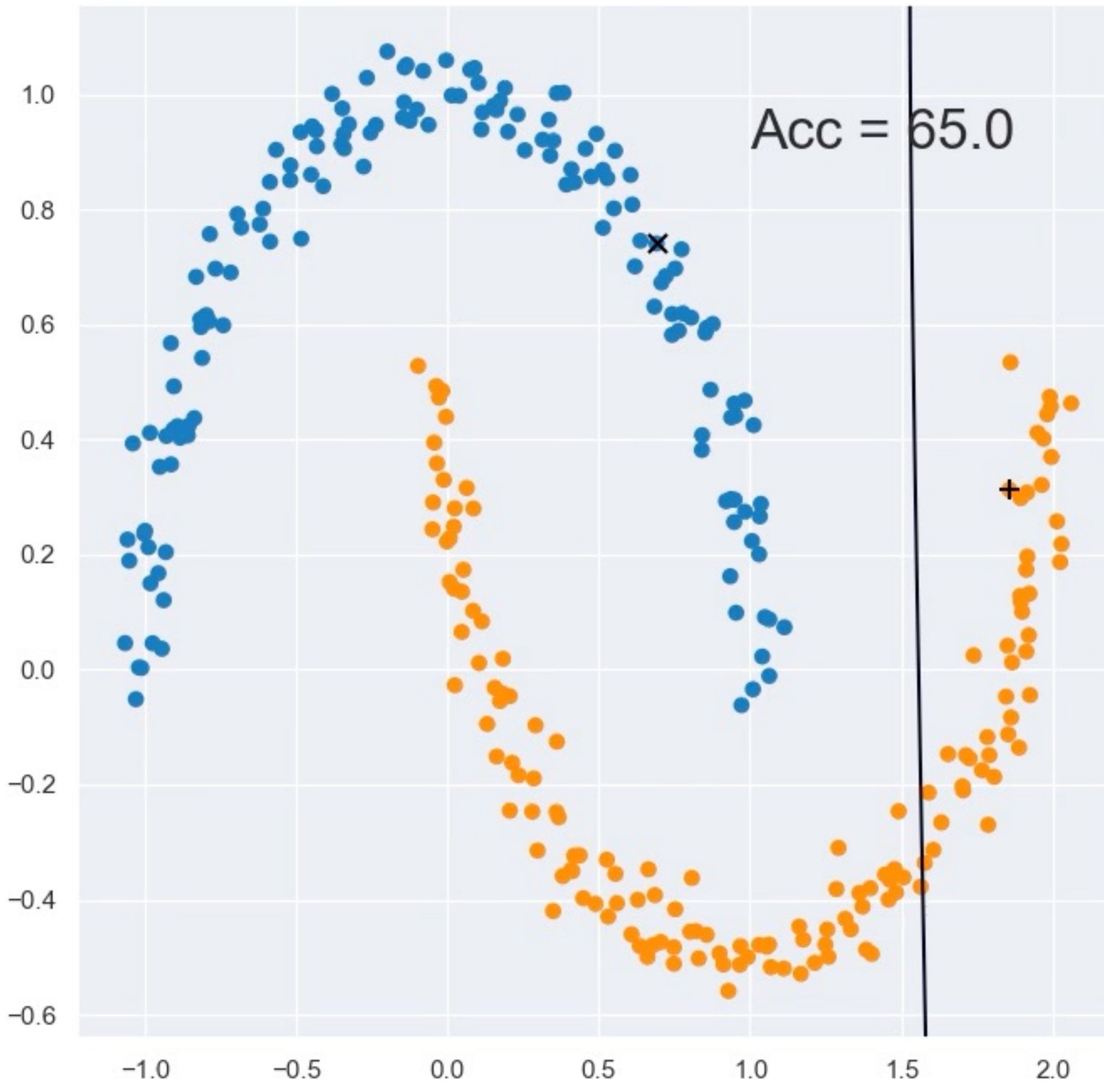


https://link.springer.com/content/pdf/10.1007/s10994-019-05855-6.pdf

Semi-Supervised Learning

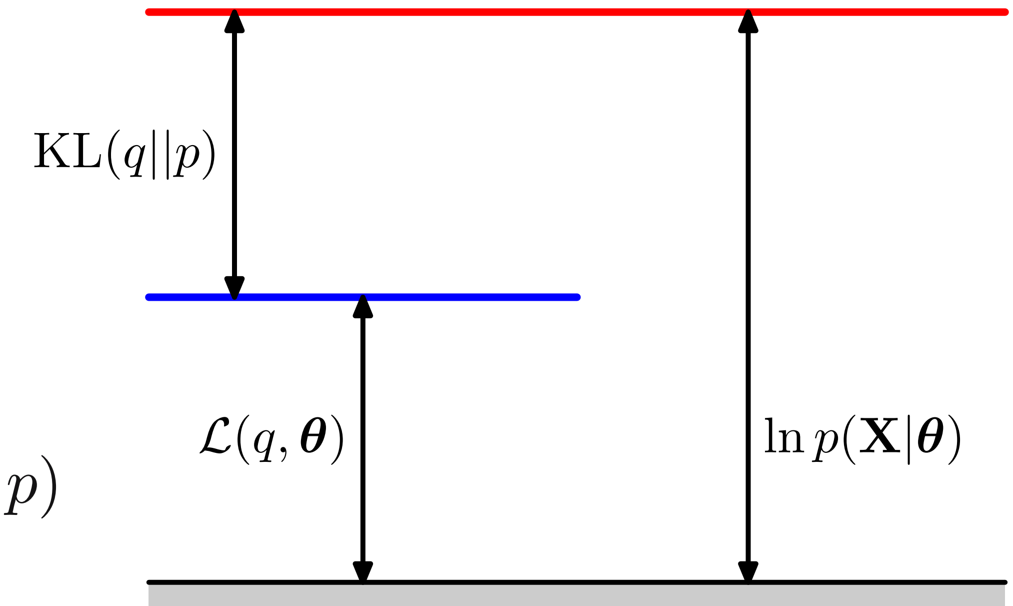


Semi-Supervised Learning



Bonus Math: EM in General

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



$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p)$$

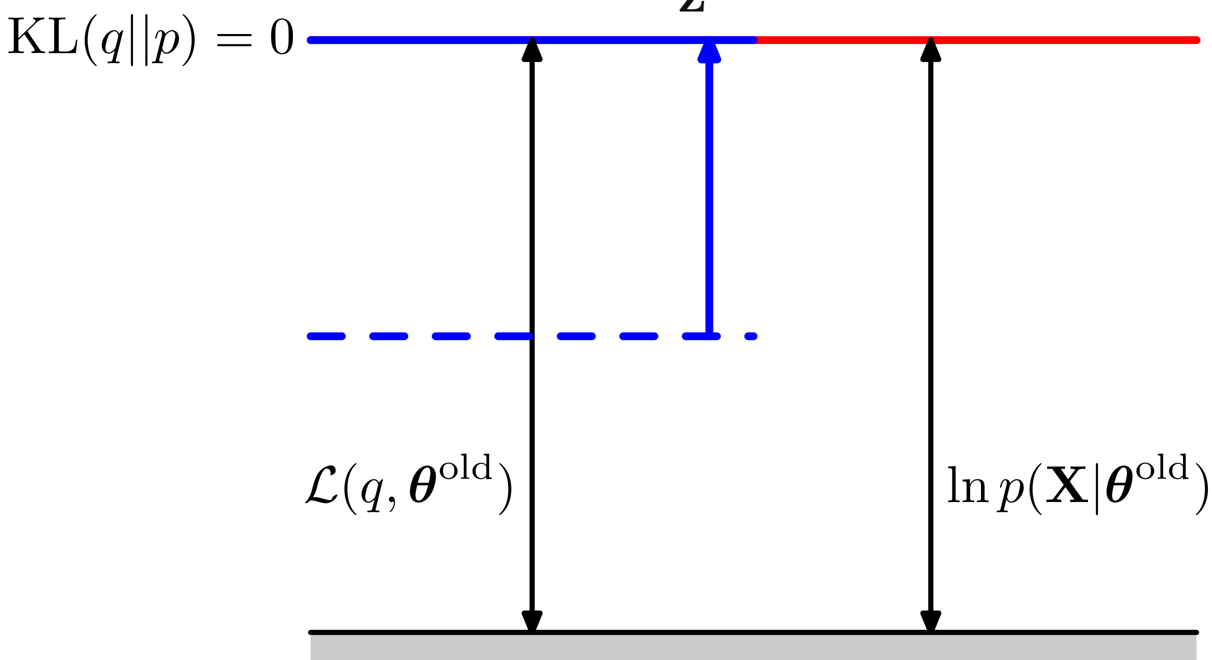
$$\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \\
\mathrm{KL}(q\|p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}
\end{aligned}$$

Pattern Recognition and Machine Learning: http://www.rmki.kfki.hu/~banmi/elte/bishop_em.pdf

EM: Pictorial View

$$\mathrm{KL}(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

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

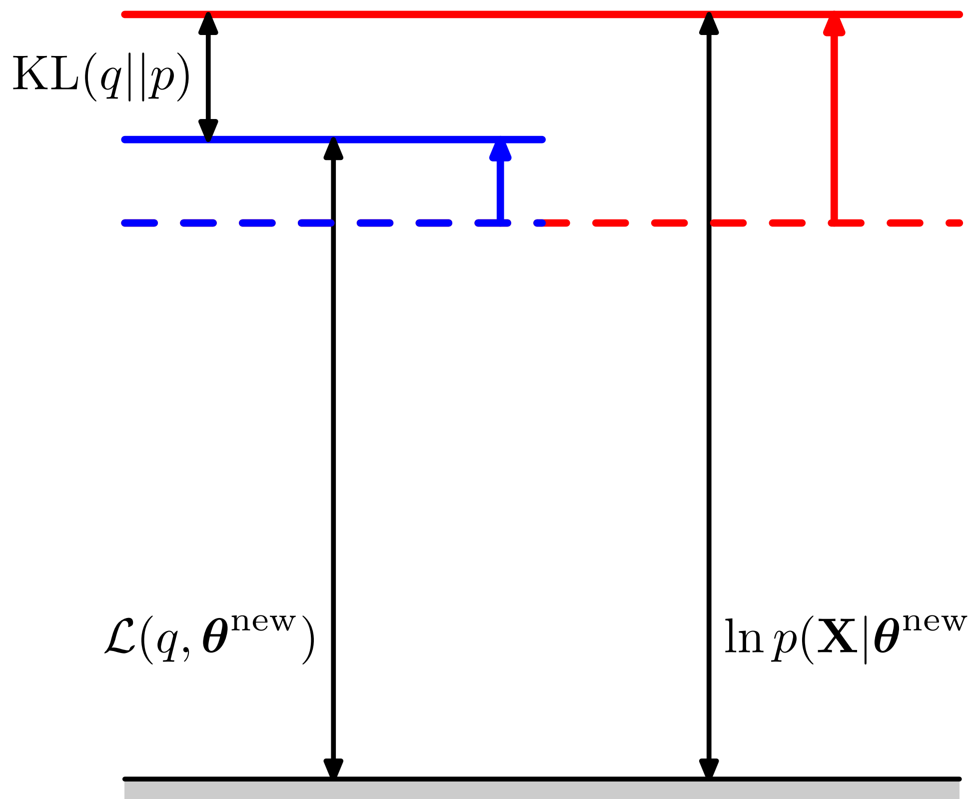


$$\begin{aligned}
\mathcal{L}(q,\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\mathrm{old}}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\mathrm{old}}) \ln p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\mathrm{old}}) \\
&= \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathrm{old}}) + \mathrm{const} \qquad (9.74)
\end{aligned}$$

Pattern Recognition and Machine Learning: http://www.rmki.kfki.hu/~banmi/elte/bishop_em.pdf

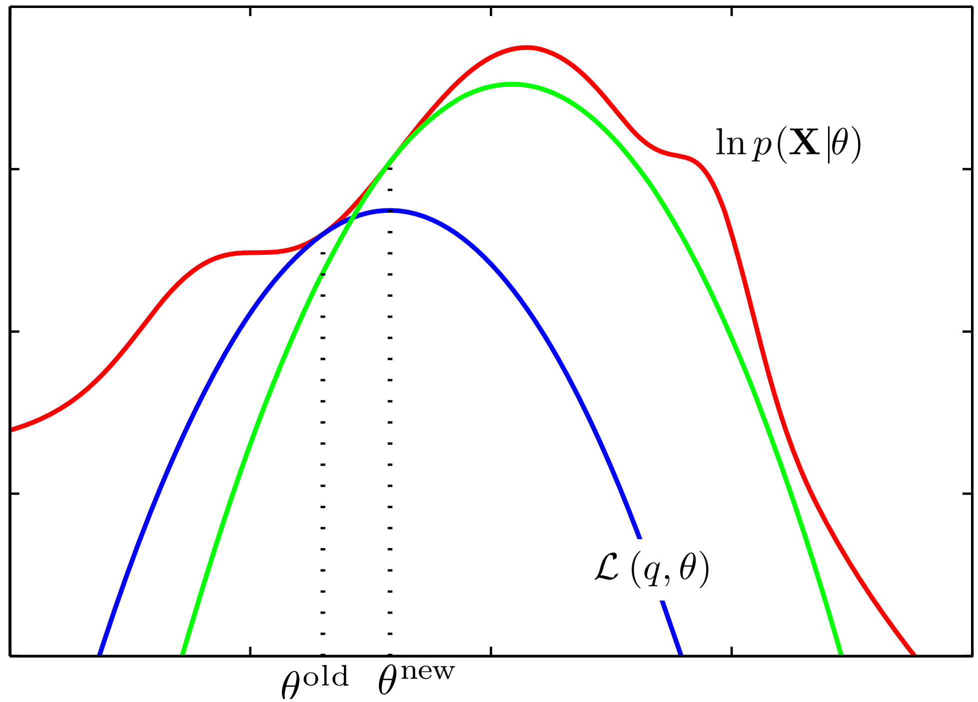
EM: Pictorial View

Illustration of the M step of the EM algorithm. The distribution $q(\mathbf{Z})$ is held fixed and the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to the parameter vector $\boldsymbol{\theta}$ to give a revised value $\boldsymbol{\theta}^{\text{new}}$. Because the KL divergence is nonnegative, this causes the log likelihood $\ln p(\mathbf{X}|\boldsymbol{\theta})$ to increase by at least as much as the lower bound does.



Pattern Recognition and Machine Learning: http://www.rmki.kfki.hu/~banmi/elte/bishop_em.pdf

EM: Pictorial View



$$\log p(X \mid \theta) = \boxed{L(q,\theta)} + \boxed{KL(q \,\|\, p)}$$

Increases   Can only increase

$$\log p(X \mid \theta) \geq \log p(X \mid \theta^{old})$$

Pattern Recognition and Machine Learning: http://www.rmki.kfki.hu/~banmi/elte/bishop_em.pdf