ECE521 tutorial
Latent Variable Models and Expectation Maximization

Presented by Renjie Liao and Eleni Triantafillou
Most slides borrowed from Jimmy Ba :)

Prior, posterior and likelihood functions

- When designing machine learning models with the parameters θ , we assume uncertainties in the parameters:
  - Prior: $P(\theta)$ (the model of engineering knowledge)
  - Likelihood: $P(\mathrm{data} \,|\, \theta)$ (the model of data)

- Learning is about finding a “good” set of parameters under our modelling assumption
  - One common approach is Maximum-likelihood estimation (MLE)

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \left[-\log P(\mathrm{data} \,|\, \theta) \right]$$

Prior, posterior and likelihood functions

- When designing machine learning models with the parameters θ , we assume uncertainties in the parameters:
  - Prior: $P(\theta)$ (the model of engineering knowledge)
  - Likelihood: $P(\mathrm{data} \,|\, \theta)$ (the model of data)
- Learning is about finding a “good” set of parameters under our modelling assumption
  - One common approach is Maximum-likelihood estimation (MLE)

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \log P(\mathrm{data} \,|\, \theta)$$

Prior, posterior and likelihood functions

- When designing machine learning models with the parameters θ , we assume uncertainties in the parameters:
    - Prior: $P(\theta)$ (the model of engineering knowledge)
    - Likelihood: $P(\mathrm{data} \mid \theta)$ (the model of data)
- Learning is about finding a “good” set of parameters under our modelling assumption
    - Another approach is Maximum a posteriori estimation (MAP)

Prior, posterior and likelihood functions

- When designing machine learning models with the parameters θ , we assume uncertainties in the parameters:
  - Prior: $P(\theta)$ (the model of engineering knowledge)
  - Likelihood: $P(\mathrm{data} \,|\, \theta)$ (the model of data)
- Learning is about finding a “good” set of parameters under our modelling assumption
  - Another approach is Maximum a posteriori estimation (MAP)

$$P(\theta \,|\, \mathrm{data}) = \frac{P(\mathrm{data} \,|\, \theta) P(\theta)}{P(\mathrm{data})}$$

Prior, posterior and likelihood functions

- When designing machine learning models with the parameters θ , we assume uncertainties in the parameters:
  - Prior: $P(\theta)$ (the model of engineering knowledge)
  - Likelihood: $P(\mathrm{data}\,|\,\theta)$ (the model of data)

- Learning is about finding a “good” set of parameters under our modelling assumption
  - Another approach is Maximum a posteriori estimation (MAP)

$$
\begin{aligned}
P(\theta\,|\,\mathrm{data}) &= \frac{P(\mathrm{data}\,|\,\theta)P(\theta)}{P(\mathrm{data})} \\
&\propto P(\mathrm{data}\,|\,\theta)P(\theta)
\end{aligned}
$$

Prior, posterior and likelihood functions

- When designing machine learning models with the parameters θ , we assume uncertainties in the parameters:
  - Prior: $P(\theta)$ (the model of engineering knowledge)
  - Likelihood: $P(\mathrm{data} \,|\, \theta)$ (the model of data)

- Learning is about finding a “good” set of parameters under our modelling assumption
  - Another approach is Maximum a posteriori estimation (MAP)

$$\theta^* = \operatorname*{argmax}_{\theta} \log P(\theta \,|\, \mathrm{data})$$

Prior, posterior and likelihood functions

- When designing machine learning models with the parameters θ , we assume uncertainties in the parameters:
  - Prior: $P(\theta)$ (the model of engineering knowledge)
  - Likelihood: $P(\mathrm{data}\,|\,\theta)$ (the model of data)

- Learning is about finding a "good" set of parameters under our modelling assumption
  - Another approach is Maximum a posteriori estimation (MAP)

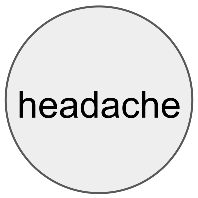
$$\begin{aligned}\theta^* &= \underset{\theta}{\operatorname{argmax}} \log P(\theta\,|\,\mathrm{data}) \\ &= \underset{\theta}{\operatorname{argmax}} \left[\log P(\mathrm{data}\,|\,\theta) + \log P(\theta)\right]\end{aligned}$$

Prior, posterior and likelihood functions

- Often, instead of learning the parameters, we would like to infer the hidden causality for the observed data

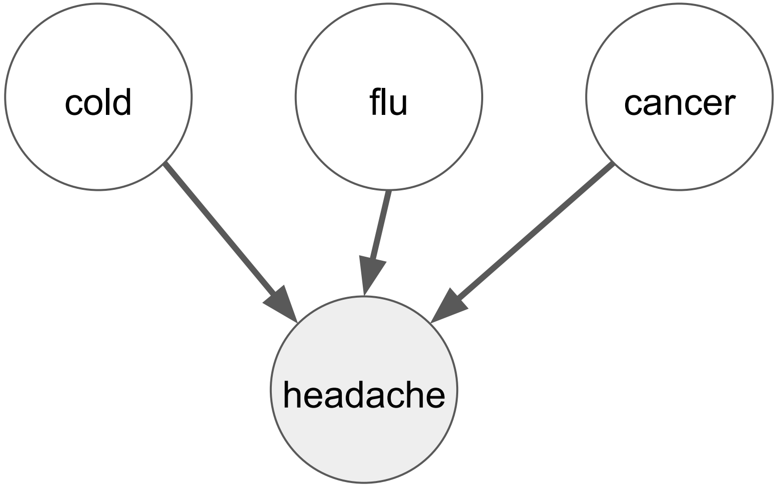
Prior, posterior and likelihood functions

- Often, instead of learning the parameters, we would like to infer the hidden causality for the observed data



Prior, posterior and likelihood functions

- Often, instead of learning the parameters, we would like to infer the hidden causality for the observed data



Prior, posterior and likelihood functions

- Build more complicated probabilistic models by introducing latent random variables z
  - Prior: $P(z)$ (the model of the world)
  - Likelihood: $P(\mathrm{data} \mid z)$ (the model of data)

Prior, posterior and likelihood functions

- Build more complicated probabilistic models by introducing latent random variables z
  - Prior: $P(z)$ (the model of the world)
  - Likelihood: $P(\mathrm{data} \mid z)$(the model of data)
- Inference is to compute the posterior distributions of the latent RVs

$$P(z \mid \mathrm{data}) = \frac{P(\mathrm{data} \mid z) P(z)}{P(\mathrm{data})}$$

Prior, posterior and likelihood functions

- Build more complicated probabilistic models by introducing latent random variables z
  - Prior: $P(z)$ (the model of the world)
  - Likelihood: $P(\mathrm{data} \mid z)$ (the model of data)

- Inference is to compute the posterior distributions of the latent RVs

$$P(z \mid \mathrm{data}) = \frac{P(\mathrm{data} \mid z) P(z)}{P(\mathrm{data})}$$

Mint factory A Mint factory B

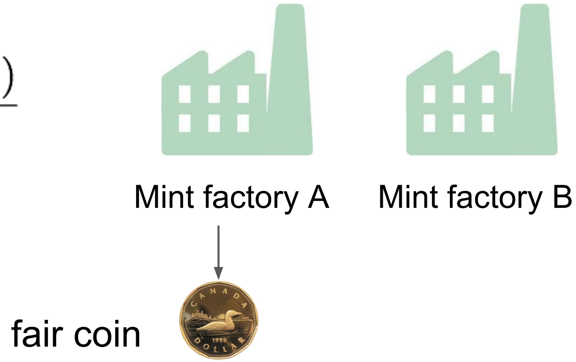
Examples:

Prior, posterior and likelihood functions

- Build more complicated probabilistic models by introducing latent random variables z
  - Prior: $P(z)$ (the model of the world)
  - Likelihood: $P(\mathrm{data} \,|\, z)$(the model of data)
- Inference is to compute the posterior distributions of the latent RVs

$$P(z \,|\, \mathrm{data}) = \frac{P(\mathrm{data} \,|\, z) P(z)}{P(\mathrm{data})}$$

Examples:

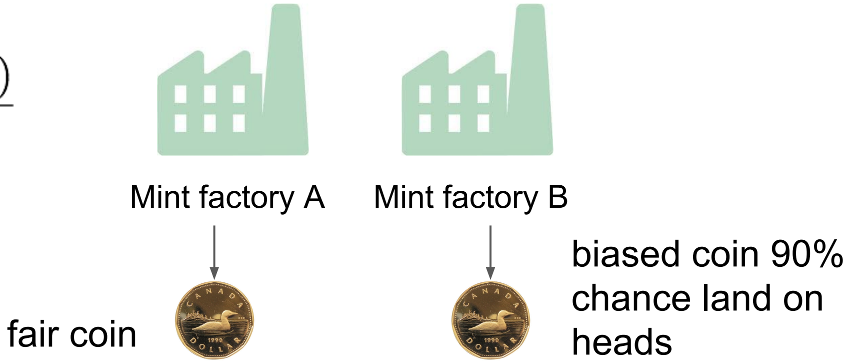


Prior, posterior and likelihood functions

- Build more complicated probabilistic models by introducing latent random variables z
  - Prior: $P(z)$ (the model of the world)
  - Likelihood: $P(\mathrm{data} \mid z)$(the model of data)
- Inference is to compute the posterior distributions of the latent RVs

$$P(z \mid \mathrm{data}) = \frac{P(\mathrm{data} \mid z)P(z)}{P(\mathrm{data})}$$

Examples:

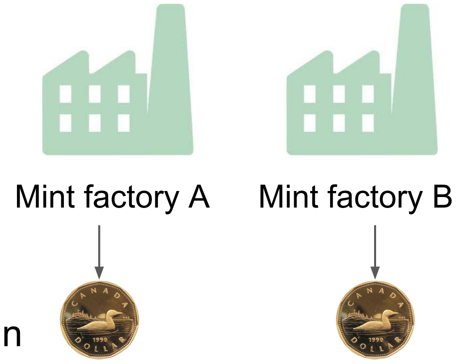


Prior, posterior and likelihood functions

- Build more complicated probabilistic models by introducing latent random variables z
  - Prior: $P(z)$ (the model of the world)
  - Likelihood: $P(\text{data} \mid z)$(the model of data)
- Inference is to compute the posterior distributions of the latent RVs

$$P(z \mid \text{data}) = \frac{P(\text{data} \mid z)P(z)}{P(\text{data})}$$

Examples:



fair coin

biased coin 90
chance land o
heads

Observe 3 coins and 4 tosses each

0	1	0	1
1	1	1	0
1	0	1	1

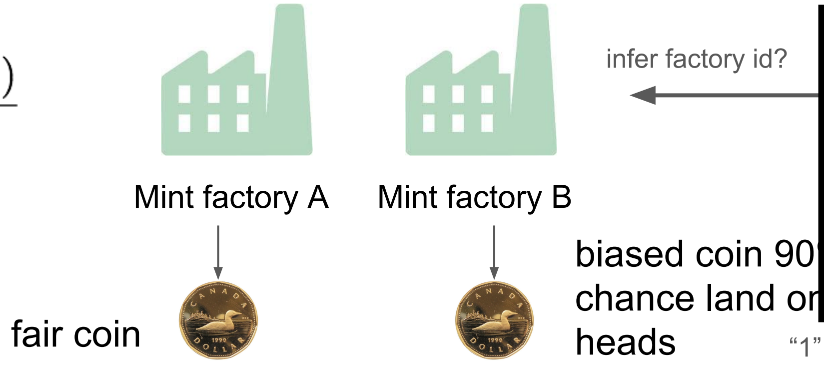
“1” denote heads, “0” denote tails

Prior, posterior and likelihood functions

- Build more complicated probabilistic models by introducing latent random variables z
  - Prior: $P(z)$ (the model of the world)
  - Likelihood: $P(\mathrm{data}\,|\,z)$(the model of data)
- Inference is to compute the posterior distributions of the latent RVs

$$P(z\,|\,\mathrm{data}) = \frac{P(\mathrm{data}\,|\,z)P(z)}{P(\mathrm{data})}$$

Examples:

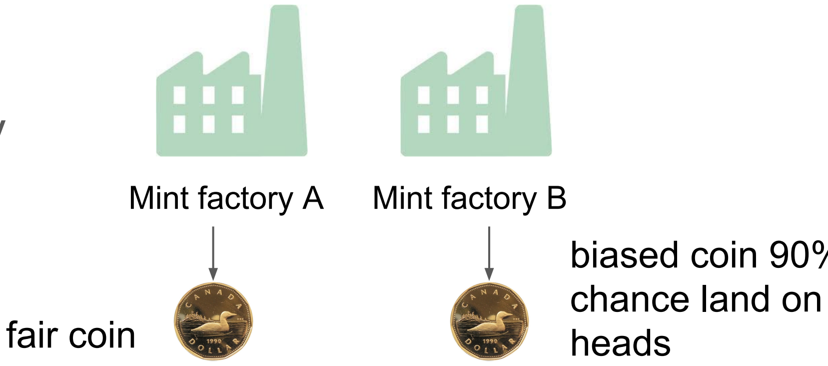


0	1	0	1
1	1	1	0
1	0	1	1

“1” denote heads, “0” denote tails

Prior, posterior and likelihood functions

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1) = P(z=2) = 0.5$
  - Likelihood: $P(x = \text{heads} \,|\, z = 1) = 0.5$
    $P(x = \text{heads} \,|\, z = 2) = 0.9$



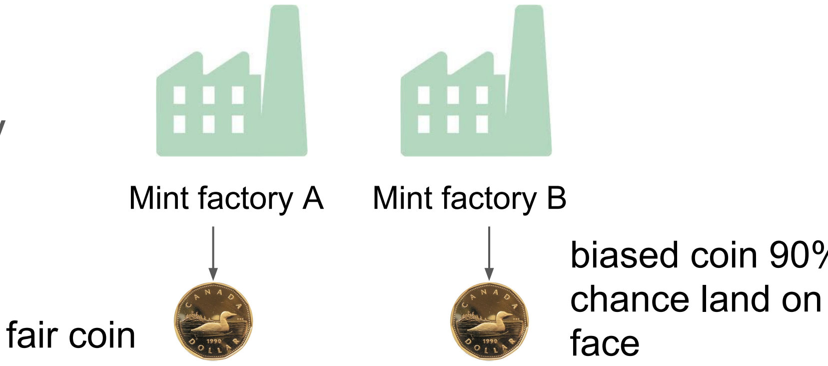
"1" denote heads, "0" denote tails

0	1	0	1
1	1	1	0
1	0	1	1

Posterior: $P(z \,|\, \text{data}) = \dfrac{P(\text{data} \,|\, z) P(z)}{P(\text{data})}$

Prior, posterior and likelihood functions

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1) = P(z=2) = 0.5$
  - Likelihood: $P(x = \text{heads} \,|\, z=1) = 0.5$
    $P(x = \text{heads} \,|\, z=2) = 0.9$



$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

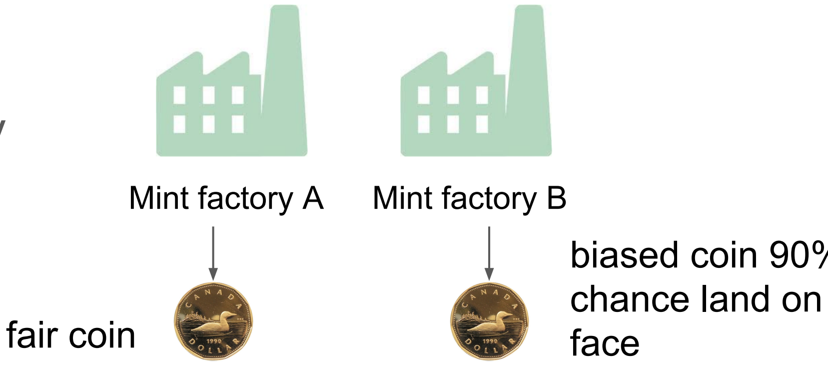
Posterior: $P(z=1 \,|\, x_1 = [0,1,0,1]) = \dfrac{P(x_1 = [0,1,0,1] \,|\, z=1)P(z=1)}{P(x_1 = [0,1,0,1])}$

$$\propto P(x=0 \,|\, z=1)^2 P(x=1 \,|\, z=1)^2 P(z=1)$$

$$= 0.5^4 * 0.5 = 0.0315$$

Prior, posterior and likelihood functions

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1) = P(z=2) = 0.5$
  - Likelihood: $P(x = \text{heads} \,|\, z=1) = 0.5$
    $P(x = \text{heads} \,|\, z=2) = 0.9$



$$P(z=1 \,|\, x_1 = [0,1,0,1]) \propto \; 0.0315$$

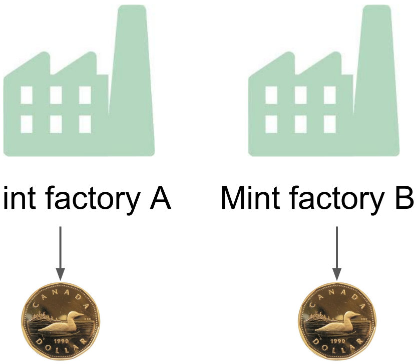
Posterior:

$$\begin{aligned} P(z=2 \,|\, x_1 = [0,1,0,1]) &= \frac{P(x_1 = [0,1,0,1] \,|\, z=2) P(z=2)}{P(x_1 = [0,1,0,1])} \\ &\propto P(x=0 \,|\, z=2)^2 P(x=1 \,|\, z=2)^2 P(z=2) \\ &= 0.9^2 * 0.1^2 * 0.5 = 0.00405 \end{aligned}$$

$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

Prior, posterior and likelihood functions

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1) = P(z=2) = 0.5$
  - Likelihood: $P(x = \text{heads} \,|\, z=1) = 0.5$
    $P(x = \text{heads} \,|\, z=2) = 0.9$



Posterior:

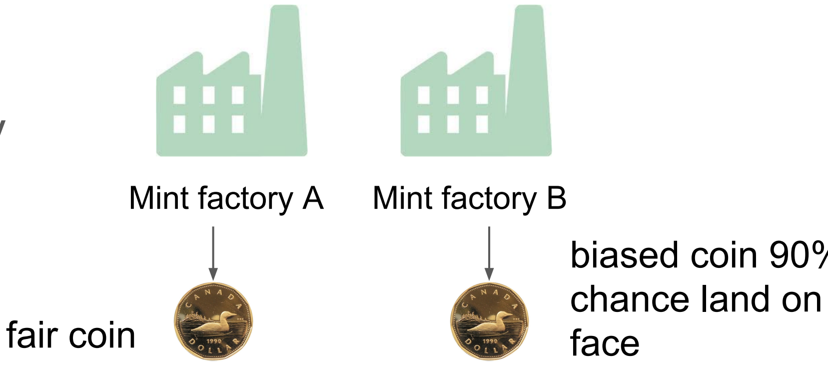
$$P(z=1 \,|\, x_1 = [0,1,0,1]) \propto 0.0315$$

$$P(z=2 \,|\, x_1 = [0,1,0,1]) \propto 0.00405$$

0	1	0	1
1	1	1	0
1	0	1	1

Prior, posterior and likelihood functions

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1) = P(z=2) = 0.5$
  - Likelihood: $P(x = \text{heads} \,|\, z=1) = 0.5$
    $P(x = \text{heads} \,|\, z=2) = 0.9$



Posterior:

$$P(z=1 \,|\, x_1 = [0,1,0,1]) \propto 0.0315$$

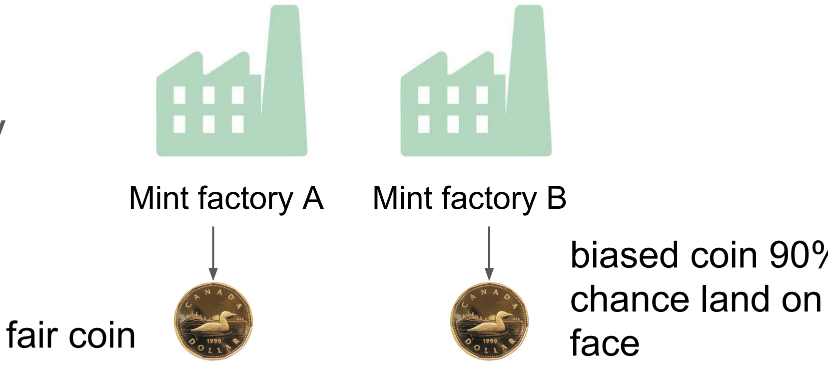
$$P(z=2 \,|\, x_1 = [0,1,0,1]) \propto 0.00405$$

$$P(z=1 \,|\, x_1 = [0,1,0,1]) = \frac{0.0315}{0.0315 + 0.00405} = 0.9$$

$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

Prior, posterior and likelihood functions

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1) = P(z=2) = 0.5$
  - Likelihood: $P(x = \text{heads} \,|\, z=1) = 0.5$
    $P(x = \text{heads} \,|\, z=2) = 0.9$



Posterior: $P(z=1 \,|\, x_1 = [0,1,0,1]) = \dfrac{0.0315}{0.0315 + 0.00405} = 0.9$

$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

How about learning the parameters for the latent variable models?

Expectation maximization

- Consider the Maximum-likelihood Estimation (MLE) approach to learn model parameters $\theta = \{\theta_{\text{prior}}, \theta_{\text{likelihood}}\}$:
  - Prior: $P(z \mid \theta_{\text{prior}})$ (the model of the world)
  - Likelihood: $P(x \mid z, \theta_{\text{likelihood}})$ (the model of data)

Expectation maximization

- Consider the Maximum-likelihood Estimation (MLE) approach to learn model parameters $\theta = \{\theta_{\text{prior}}, \theta_{\text{likelihood}}\}$:
  - Prior: $P(z \mid \theta_{\text{prior}})$ (the model of the world), e.g. mint example
  - Likelihood: $P(x \mid z, \theta_{\text{likelihood}})$ (the model of data)

$$P(z=1) = \theta_{\text{prior}}$$
$$P(z=2) = 1 - \theta_{\text{prior}}$$
$$P(x = \text{heads} \mid z=1) = \theta_{\text{likelihood}1}$$
$$P(x = \text{heads} \mid z=2) = \theta_{\text{likelihood}2}$$

Expectation maximization

- Consider the Maximum-likelihood Estimation (MLE) approach to learn model parameters $\theta = \{\theta_{\text{prior}}, \theta_{\text{likelihood}}\}$:
    - Prior: $P(z \,|\, \theta_{\text{prior}})$ (the model of the world)
    - Likelihood: $P(x \,|\, z, \theta_{\text{likelihood}})$ (the model of data)
- Define data likelihood or marginal likelihood as:

$$P(x \,|\, \theta) = \sum_z P(z \,|\, \theta_{\text{prior}}) P(x \,|\, z, \theta_{\text{likelihood}})$$

Expectation maximization

- Consider the Maximum-likelihood Estimation (MLE) approach to learn model parameters $\theta = \{\theta_{\text{prior}}, \theta_{\text{likelihood}}\}$:
  - Prior: $P(z \mid \theta_{\text{prior}})$ (the model of the world)
  - Likelihood: $P(x \mid z, \theta_{\text{likelihood}})$ (the model of data)

- Define data likelihood or marginal likelihood as:

$$P(x \mid \theta) = \sum_z P(z \mid \theta_{\text{prior}}) P(x \mid z, \theta_{\text{likelihood}})$$

- MLE of the model parameter:

$$\theta^* = \operatorname*{argmax}_\theta \log P(x \mid \theta) = \operatorname*{argmax}_\theta \log \sum_z P(z \mid \theta) P(x \mid z, \theta)$$

Learning: Expectation maximization

- Main idea: maximize a “tight” lower bound will also improve the marginal log likelihood

$$\log P(x \mid \theta) = \log \sum_z P(z \mid \theta) P(x \mid z, \theta)$$

Learning: Expectation maximization

- Main idea: maximize a “tight” lower bound will also improve the marginal log likelihood

$$\log P(x \mid \theta) = \log \sum_z P(z \mid \theta) P(x \mid z, \theta)$$

$$\log \sum_z P(z \mid \theta) P(x \mid z, \theta) = \log \sum_z Q(z) P(z \mid \theta) P(x \mid z, \theta) / Q(z)$$

Learning: Expectation maximization

- Main idea: maximize a “tight” lower bound will also improve the marginal log likelihood

$$\log P(x \mid \theta) = \log \sum_z P(z \mid \theta) P(x \mid z, \theta)$$

$$\begin{aligned}\log \sum_z P(z \mid \theta) P(x \mid z, \theta) &= \log \sum_z Q(z) P(z \mid \theta) P(x \mid z, \theta) / Q(z) \\ &= \log \mathbb{E}_{Q(z)} \left[P(z \mid \theta) P(x \mid z, \theta) / Q(z) \right]\end{aligned}$$

Learning: Expectation maximization

- Main idea: maximize a “tight” lower bound will also improve the marginal log likelihood

$$\log P(x \mid \theta) = \log \sum_z P(z \mid \theta) P(x \mid z, \theta)$$

$$\begin{aligned}
\log \sum_z P(z \mid \theta) P(x \mid z, \theta) &= \log \sum_z Q(z) P(z \mid \theta) P(x \mid z, \theta) / Q(z) \\
&= \log \mathbb{E}_{Q(z)} \left[P(z \mid \theta) P(x \mid z, \theta) / Q(z) \right] \\
&\geq \mathbb{E}_{Q(z)} \left[\log \frac{P(z \mid \theta) P(x \mid z, \theta)}{Q(z)} \right]
\end{aligned}$$

Jensen’s Inequality

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

f() is log that is concave

Learning: Expectation maximization

- Main idea: maximize a “tight” lower bound will also improve the marginal log likelihood

$$\log P(x \,|\, \theta) = \log \sum_z P(z \,|\, \theta) P(x \,|\, z, \theta)$$

$$\begin{aligned}
\log \sum_z P(z \,|\, \theta) P(x \,|\, z, \theta) &= \log \sum_z Q(z) P(z \,|\, \theta) P(x \,|\, z, \theta) / Q(z) \\
&= \log \mathbb{E}_{Q(z)} \left[P(z \,|\, \theta) P(x \,|\, z, \theta) / Q(z) \right] \\
&\geq \mathbb{E}_{Q(z)} \left[\log \frac{P(z \,|\, \theta) P(x \,|\, z, \theta)}{Q(z)} \right] \\
&= \sum_z Q(z) \log \frac{P(z \,|\, \theta) P(x \,|\, z, \theta)}{Q(z)}
\end{aligned}$$

Learning: Expectation maximization

- Main idea: maximize a "tight" lower bound will also improve the marginal log likelihood

$$\log \sum_z P(z \mid \theta) P(x \mid z, \theta) = \log \sum_z Q(z) P(z \mid \theta) P(x \mid z, \theta) / Q(z)$$

lower bound:

$$\geq \sum_z Q(z) \log \frac{P(z \mid \theta) P(x \mid z, \theta)}{Q(z)}$$

- First, we ensure the lower bound is tight to the marginal log likelihood
  - Find the Q distribution for which the equality holds in the Jensen's Inequality:

tighten the lower bound:

$$Q(z) \propto P(z \mid \theta) P(x \mid z, \theta)$$

$$\text{i.e.} Q(z) = P(z \mid x, \theta)$$

- Second, optimize the parameters in the lower bound

optimize the lower bound:

$$\theta^* = \operatorname*{argmax}_\theta \sum_z P(z \mid x) \log P(z \mid \theta) P(x \mid z, \theta)$$

Learning: Expectation maximization

- Main idea: maximize a “tight” lower bound will also improve the marginal log likelihood

$$\log \sum_z P(z \mid \theta) P(x \mid z, \theta) = \log \sum_z Q(z) P(z \mid \theta) P(x \mid z, \theta) / Q(z)$$

lower bound:

$$\geq \sum_z Q(z) \log \frac{P(z \mid \theta) P(x \mid z, \theta)}{Q(z)}$$

- First, we ensure the lower bound is tight to the marginal log likelihood
  - Find the Q distribution for which the equality holds in the Jensen’s Inequality:

tighten the lower bound:

$$Q(z) \propto P(z \mid \theta) P(x \mid z, \theta)$$

$$\text{i.e.} Q(z) = P(z \mid x, \theta)$$

- Second, optimize the parameters in the lower bound

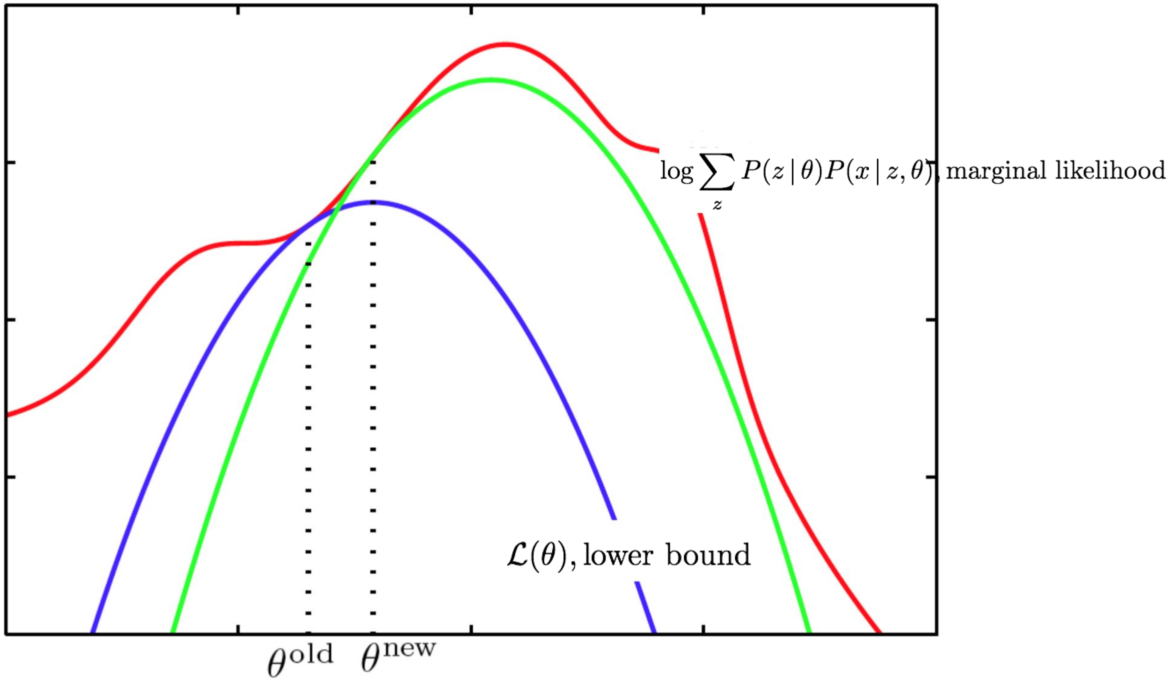
optimize the lower bound:

$$\theta^* = \operatorname*{argmax}_\theta \sum_z P(z \mid x) \log P(z \mid \theta) P(x \mid z, \theta)$$

repeat till convergence

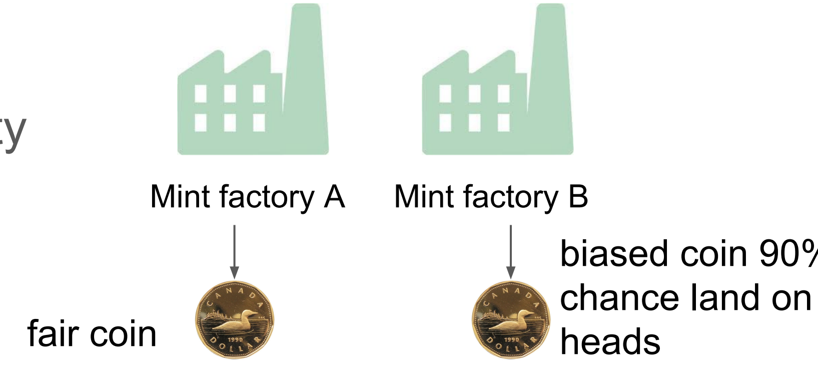
Expectation maximization

- Two steps of the EM algorithm:



Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1) = P(z=2) = 0.5$
  - Likelihood: $P(x = \text{heads} \,|\, z=1) = 0.5$
    $P(x = \text{heads} \,|\, z=2) = 0.9$



Posterior:

$$P(z_1 = 1 \,|\, x_2 = [0,1,0,1]) = \frac{0.0315}{0.0315 + 0.00405} = 0.9$$

$$P(z_2 = 1 \,|\, x_2 = [1,1,1,0]) = \frac{0.0315}{0.0315 + 0.03645} = 0.459$$

$$P(z_3 = 1 \,|\, x_2 = [1,0,1,1]) = \frac{0.0315}{0.0315 + 0.03645} = 0.459$$

“1” denote heads, “0” denote tail

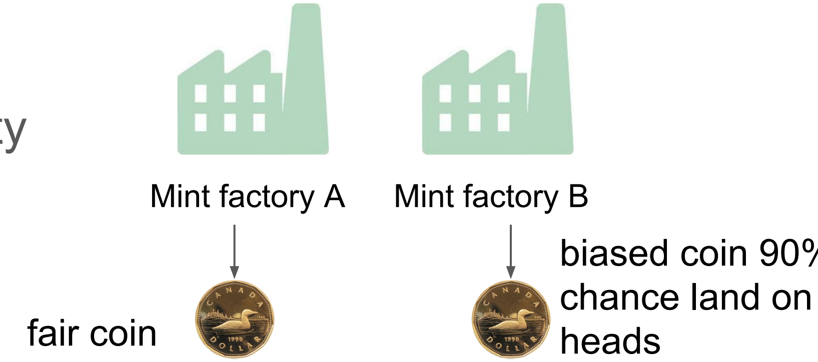
3 unknown coins, each tossed 4 times

$$\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1)=\theta_z, P(z=2)=1-\theta_z$
  - Likelihood: $P(x=\text{heads}\,|\,z=1)=\theta_1$

    $P(x=\text{heads}\,|\,z=2)=\theta_2$



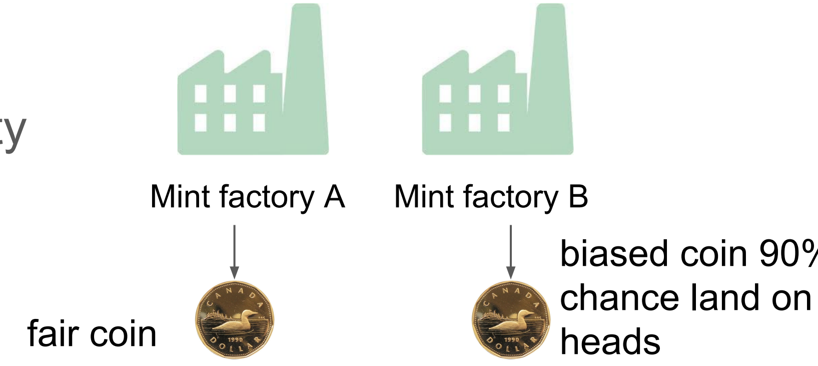
“1” denote heads, “0” denote tail

3 unknown coins, each tossed 4 times

x_1	0	1	0	1
x_2	1	1	1	0
x_3	1	0	1	1

Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1)=\theta_z, P(z=2)=1-\theta_z$
  - Likelihood: $P(x=\text{heads}\,|\,z=1)=\theta_1$
    $P(x=\text{heads}\,|\,z=2)=\theta_2$



Learn new parameters:

$$\theta_z^*, \theta_1^*, \theta_2^* = \underset{\theta_z, \theta_1, \theta_2}{\text{argmax}} \sum_n \sum_z P(z_n\,|\,x_n) \log P(z_n\,|\,\theta_z) P(x_n\,|\,z_n, \theta_1, \theta_2)$$

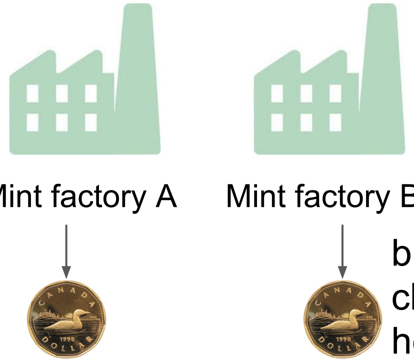
“1” denote heads, “0” denote tail

3 unknown coins, each tossed 4 times

x_1	0	1	0	1
x_2	1	1	1	0
x_3	1	0	1	1

Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1)=\theta_z, P(z=2)=1-\theta_z$
  - Likelihood: $P(x=\text{heads}\,|\,z=1)=\theta_1$
    $P(x=\text{heads}\,|\,z=2)=\theta_2$



Mint factory A

Mint factory B

fair coin

biased coin 90% chance land on heads

Learn new parameters:

$$\theta_z^*, \theta_1^*, \theta_2^* = \underset{\theta_z,\theta_1,\theta_2}{\operatorname{argmax}} \sum_n \sum_z P(z_n\,|\,x_n) \log P(z_n\,|\,\theta_z) P(x_n\,|\,z_n,\theta_1,\theta_2)$$
$$= \underset{\theta_z,\theta_1,\theta_2}{\operatorname{argmax}} \sum_n \sum_z P(z_n\,|\,x_n) [\log P(z_n\,|\,\theta_z) + \log P(x_n\,|\,z_n,\theta_1,\theta_2)]$$

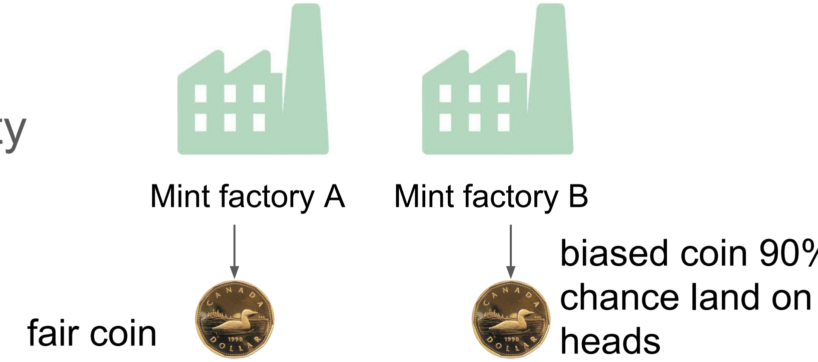
3 unknown coins, each tossed 4 times

“1” denote heads, “0” denote tail

x_1	0	1	0	1
x_2	1	1	1	0
x_3	1	0	1	1

Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1)=\theta_z, P(z=2)=1-\theta_z$
  - Likelihood: $P(x=\text{heads}\,|\,z=1)=\theta_1$
    $P(x=\text{heads}\,|\,z=2)=\theta_2$



$$\theta_z^*,\theta_1^*,\theta_2^* = \underset{\theta_z,\theta_1,\theta_2}{\operatorname{argmax}} \sum_n \sum_z P(z_n\,|\,x_n)\log P(z_n\,|\,\theta_z)P(x_n\,|\,z_n,\theta_1,\theta_2)$$

Denote:

$$\mathcal{F} = \sum_n \sum_z P(z_n\,|\,x_n)[\log P(z_n\,|\,\theta_z) + \log P(x_n\,|\,z_n,\theta_1,\theta_2)]$$

"1" denote heads, "0" denote tail

3 unknown coins, each tossed 4 times

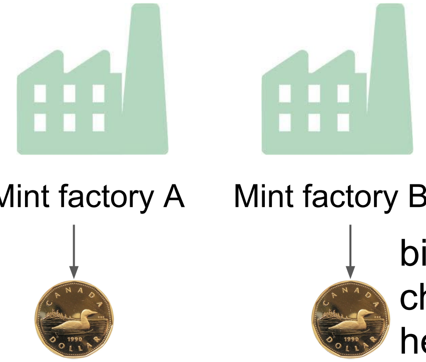
x_1	0	1	0	1
x_2	1	1	1	0
x_3	1	0	1	1

Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1)=\theta_z, P(z=2)=1-\theta_z$
  - Likelihood: $P(x=\text{heads}\,|\,z=1)=\theta_1$
    $P(x=\text{heads}\,|\,z=2)=\theta_2$

$$\mathcal{F}=\sum_n \sum_z P(z_n\,|\,x_n)[\log P(z_n\,|\,\theta_z)+\log P(x_n\,|\,z_n,\theta_1,\theta_2)]$$

Rewrite: $P(z_n|\theta_z)=\theta_z^{\{z=1\}}(1-\theta_z)^{\{z=2\}}$



“1” denote heads, “0” denote tail

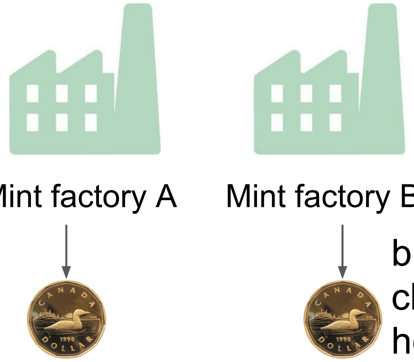
3 unknown coins, each tossed 4 times

x_1	0	1	0	1
x_2	1	1	1	0
x_3	1	0	1	1

Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1)=\theta_z, P(z=2)=1-\theta_z$
  - Likelihood: $P(x=\text{heads}\,|\,z=1)=\theta_1$

    $P(x=\text{heads}\,|\,z=2)=\theta_2$



$$\mathcal{F}=\sum_n\sum_z P(z_n\,|\,x_n)[\log P(z_n\,|\,\theta_z)+\log P(x_n\,|\,z_n,\theta_1,\theta_2)]$$

Rewrite: $$P(z_n|\theta_z)=\theta_z^{\{z=1\}}(1-\theta_z)^{\{z=2\}}$$

$$\log P(z_n|\theta_z)=\{z=1\}\log\theta_z+\{z=2\}\log(1-\theta_z)$$

“1” denote heads, “0” denote tail

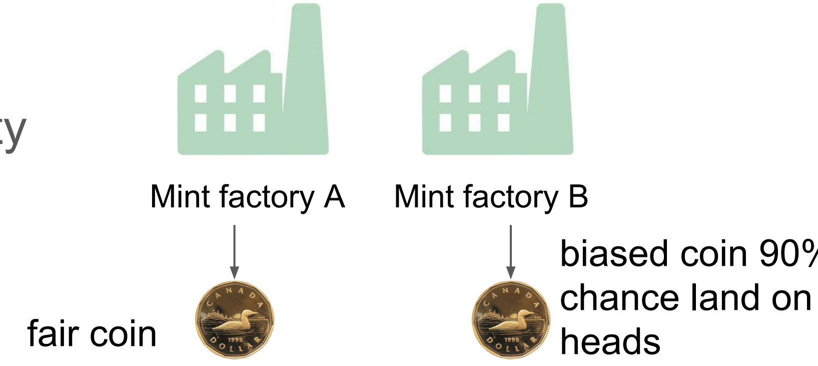
3 unknown coins, each tossed 4 times

$$\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1)=\theta_z, P(z=2)=1-\theta_z$
  - Likelihood: $P(x=\text{heads}\,|\,z=1)=\theta_1$

    $P(x=\text{heads}\,|\,z=2)=\theta_2$



$$\mathcal{F}=\sum_n\sum_z P(z_n\,|\,x_n)[\log P(z_n\,|\,\theta_z)+\log P(x_n\,|\,z_n,\theta_1,\theta_2)]$$

Rewrite: $$\frac{\partial\mathcal{F}}{\partial\theta_z}=\sum_n\sum_z P(z_n\,|\,x_n)\left[\frac{\partial}{\partial\theta_z}\log P(z_n\,|\,\theta_z)\right]$$

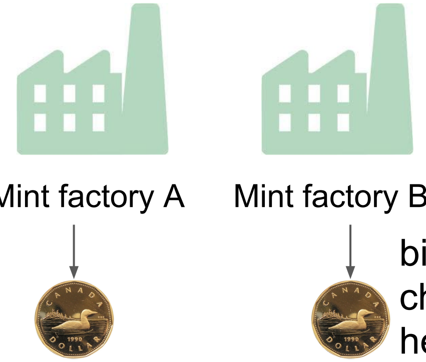
"1" denote heads, "0" denote tail

3 unknown coins, each tossed 4 times

x_1	0	1	0	1
x_2	1	1	1	0
x_3	1	0	1	1

Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1)=\theta_z, P(z=2)=1-\theta_z$
  - Likelihood: $P(x=\text{heads}\,|\,z=1)=\theta_1$
    $P(x=\text{heads}\,|\,z=2)=\theta_2$



$$\mathcal{F}=\sum_n\sum_z P(z_n\,|\,x_n)[\log P(z_n\,|\,\theta_z)+\log P(x_n\,|\,z_n,\theta_1,\theta_2)]$$

Rewrite: $$\frac{\partial\mathcal{F}}{\partial\theta_z}=\sum_n\sum_z P(z_n\,|\,x_n)\left[\frac{\partial}{\partial\theta_z}\log P(z_n\,|\,\theta_z)\right]$$

$$=0$$ set derivative to zero to solve for optimals

“1” denote heads, “0” denote tail

3 unknown coins, each tossed 4 times

x_1	0	1	0	1
x_2	1	1	1	0
x_3	1	0	1	1

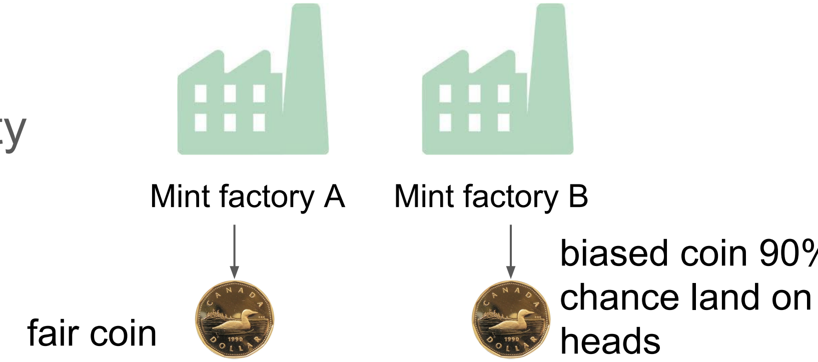
Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1)=\theta_z, P(z=2)=1-\theta_z$
  - Likelihood: $P(x=\text{heads}\,|\,z=1)=\theta_1$
    $P(x=\text{heads}\,|\,z=2)=\theta_2$

$$\frac{\partial \mathcal{F}}{\partial \theta_z}=\sum_n \sum_z P(z_n\,|\,x_n)\left[\frac{\partial}{\partial \theta_z}\log P(z_n\,|\,\theta_z)\right]$$

$$=0$$

substitute $\log P(z_n|\theta_z)=\{z=1\}\log\theta_z+\{z=2\}\log(1-\theta_z)$



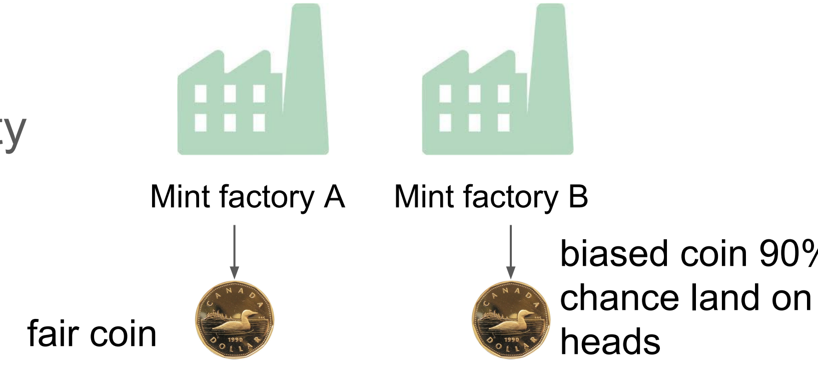
"1" denote heads, "0" denote tail

3 unknown coins, each tossed 4 times

x_1	0	1	0	1
x_2	1	1	1	0
x_3	1	0	1	1

Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1)=\theta_z, P(z=2)=1-\theta_z$
  - Likelihood: $P(x=\text{heads}\,|\,z=1)=\theta_1$
    $P(x=\text{heads}\,|\,z=2)=\theta_2$



$$\theta_z = \frac{\sum_n P(z_n=1|x_n)}{\sum_n P(z_n=1|x_n) + \sum_n P(z_n=2|x_n)}$$

“1” denote heads, “0” denote tail

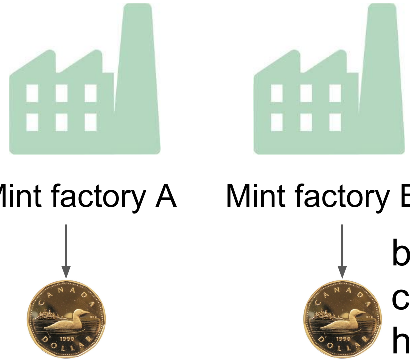
3 unknown coins, each tossed 4 times

$$\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array} \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix}$$

Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1)=\theta_z, P(z=2)=1-\theta_z$
  - Likelihood: $P(x = \text{heads} \,|\, z=1) = \theta_1$
    $P(x = \text{heads} \,|\, z=2) = \theta_2$

$$\theta_z = \frac{\sum_n P(z_n=1|x_n)}{\sum_n P(z_n=1|x_n) + \sum_n P(z_n=2|x_n)}$$
$$= \frac{\sum_n P(z_n=1|x_n)}{N}$$



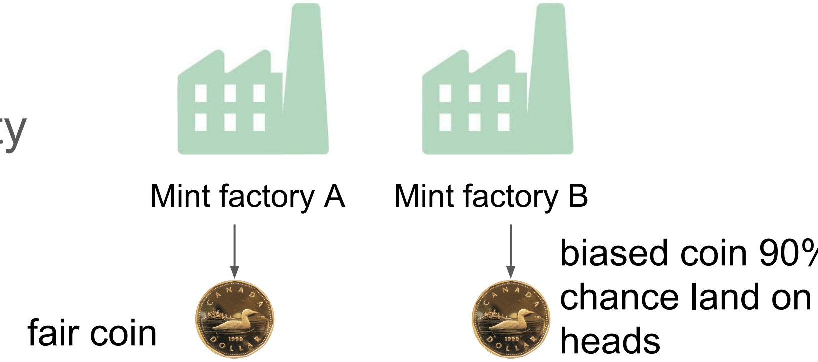
"1" denote heads, "0" denote tail

3 unknown coins, each tossed 4 times

x_1	0	1	0	1
x_2	1	1	1	0
x_3	1	0	1	1

Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1)=\theta_z, P(z=2)=1-\theta_z$
  - Likelihood: $P(x=\text{heads}\,|\,z=1)=\theta_1$
    $P(x=\text{heads}\,|\,z=2)=\theta_2$



$$
\begin{aligned}
\theta_z &= \frac{\sum_n P(z_n=1|x_n)}{\sum_n P(z_n=1|x_n)+\sum_n P(z_n=2|x_n)}\\
&= \frac{\sum_n P(z_n=1|x_n)}{N}
\end{aligned}
$$

$$P(z_1=1\,|\,x_2=[0,1,0,1]) = \frac{0.0315}{0.0315+0.00405} = 0.9$$

$$P(z_2=1\,|\,x_2=[1,1,1,0]) = \frac{0.0315}{0.0315+0.03645} = 0.459$$

$$P(z_3=1\,|\,x_2=[1,0,1,1]) = \frac{0.0315}{0.0315+0.03645} = 0.459$$

“1” denote heads, “0” denote tail

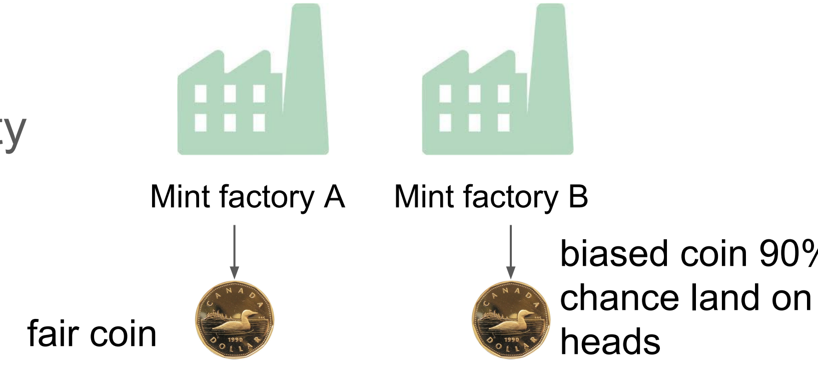
3 unknown coins, each tossed 4 times

x_1	0	1	0	1
x_2	1	1	1	0
x_3	1	0	1	1

Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1)=\theta_z, P(z=2)=1-\theta_z$
  - Likelihood: $P(x=\text{heads}\,|\,z=1)=\theta_1$

    $P(x=\text{heads}\,|\,z=2)=\theta_2$



$$\theta_z = \frac{\sum_n P(z_n=1|x_n)}{\sum_n P(z_n=1|x_n)+\sum_n P(z_n=2|x_n)}$$
$$= \frac{\sum_n P(z_n=1|x_n)}{N}$$

$$P(z_1=1\,|\,x_2=[0,1,0,1]) = \frac{0.0315}{0.0315+0.00405}=0.9$$

$$P(z_2=1\,|\,x_2=[1,1,1,0]) = \frac{0.0315}{0.0315+0.03645}=0.459$$

$$P(z_3=1\,|\,x_2=[1,0,1,1]) = \frac{0.0315}{0.0315+0.03645}=0.459$$

$$\theta_z = \frac{0.9+0.459+0.459}{3}=0.603$$

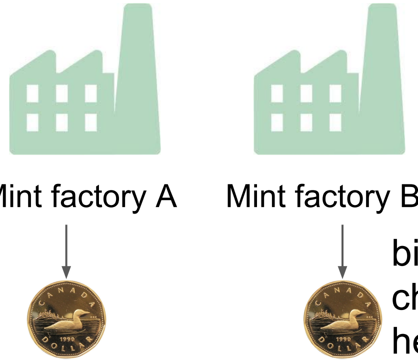
"1" denote heads, "0" denote tail

3 unknown coins, each tossed 4 times

x_1	0	1	0	1
x_2	1	1	1	0
x_3	1	0	1	1

Learning: Expectation maximization

- Mixtures of Bernoullis model
  - z = {1, 2}, denote the factory identity
  - Prior: $P(z=1)=\theta_z, P(z=2)=1-\theta_z$
  - Likelihood: $P(x=\text{heads}\,|\,z=1)=\theta_1$
    $P(x=\text{heads}\,|\,z=2)=\theta_2$



Mint factory A

Mint factory B

fair coin

biased coin 90% chance land on heads

“1” denote heads, “0” denote tail

3 unknown coins, each tossed 4 times

x_1	0	1	0	1
x_2	1	1	1	0
x_3	1	0	1	1

$$\theta_z = \frac{\sum_n P(z_n=1|x_n)}{\sum_n P(z_n=1|x_n)+\sum_n P(z_n=2|x_n)} = \frac{\sum_n P(z_n=1|x_n)}{N}$$

$$P(z_1=1\,|\,x_2=[0,1,0,1]) = \frac{0.0315}{0.0315+0.00405}=0.9$$

$$P(z_2=1\,|\,x_2=[1,1,1,0]) = \frac{0.0315}{0.0315+0.03645}=0.459$$

$$P(z_3=1\,|\,x_2=[1,0,1,1]) = \frac{0.0315}{0.0315+0.03645}=0.459$$

$$\theta_z = \frac{0.9+0.459+0.459}{3}=0.603$$

Homework question: what about θ_1, θ_2

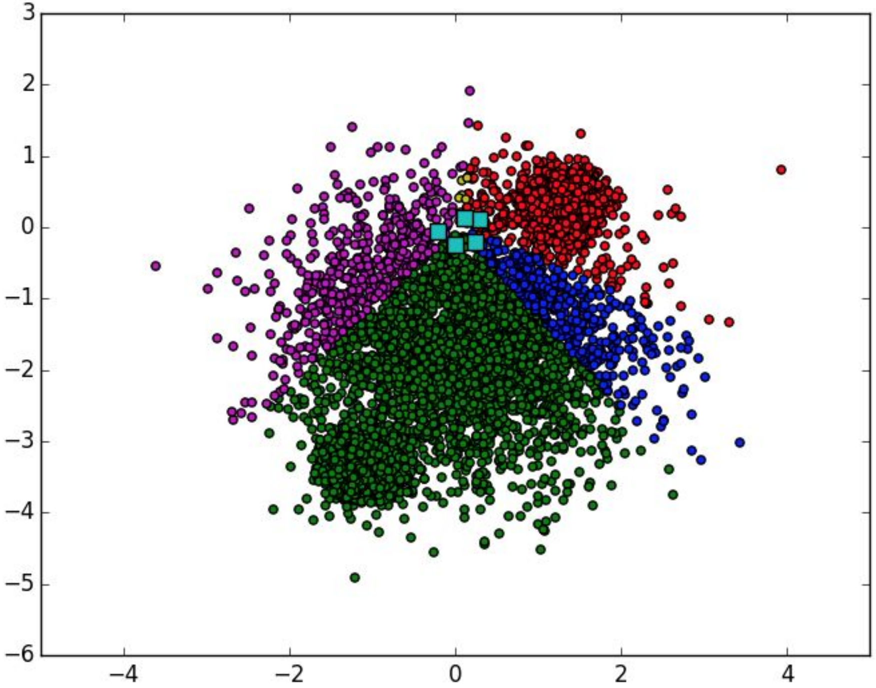
Introduction to Assignment 3

- Kmeans

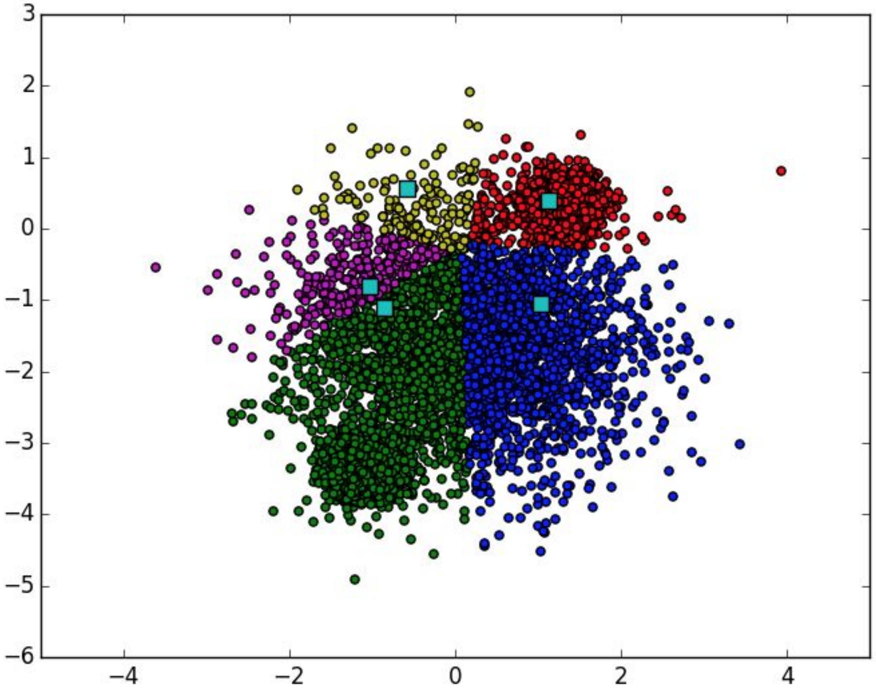
$$\mathcal{L}(\boldsymbol{\mu}) = \sum_{n=1}^{B} \min_{k=1}^{K} ||\mathbf{x}_n - \boldsymbol{\mu}_k||_2^2.$$

  - Data: $\mathbf{x}_n$
  - Cluster center: $\boldsymbol{\mu}_k$
  - Cluster assignment: $\min_{k=1}^{K} ||\mathbf{x}_n - \boldsymbol{\mu}_k||_2^2$

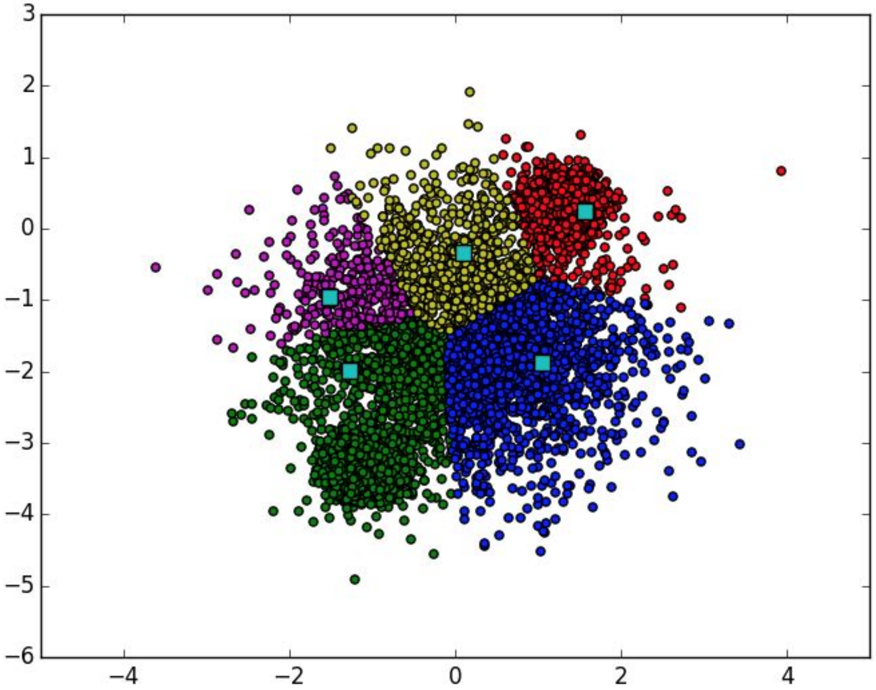
Introduction to Assignment 3



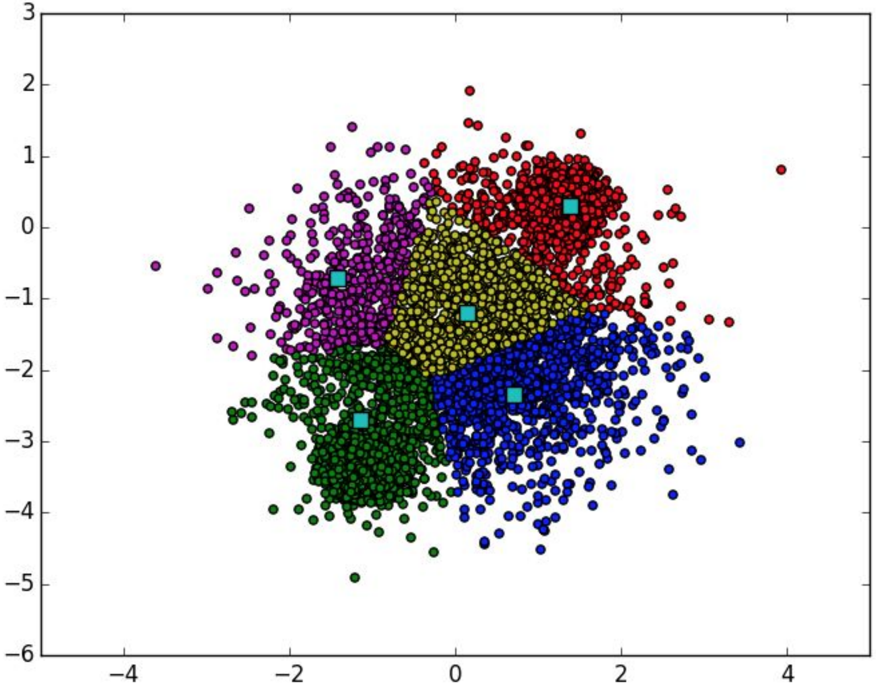
Introduction to Assignment 3



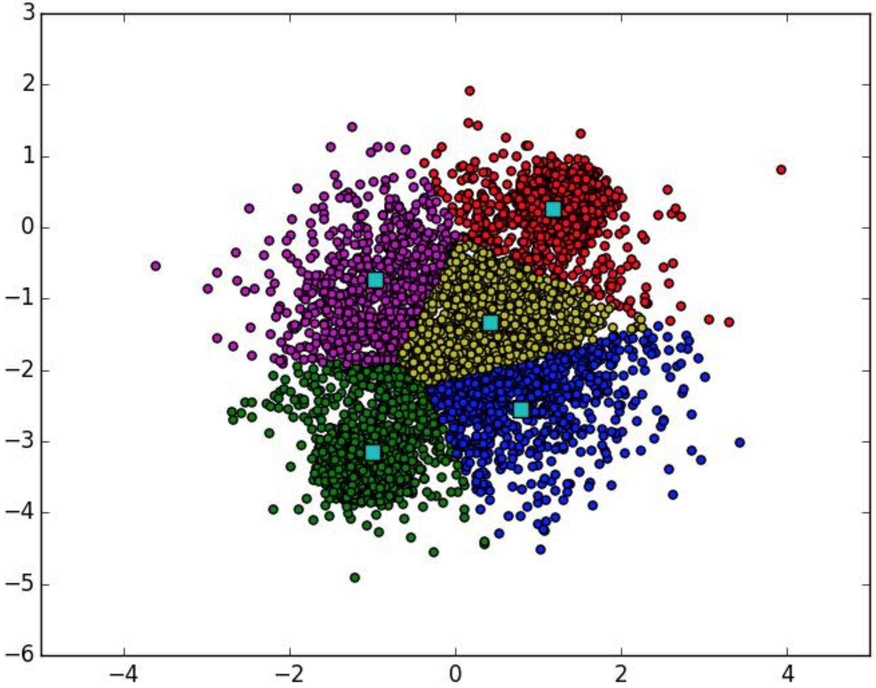
Introduction to Assignment 3



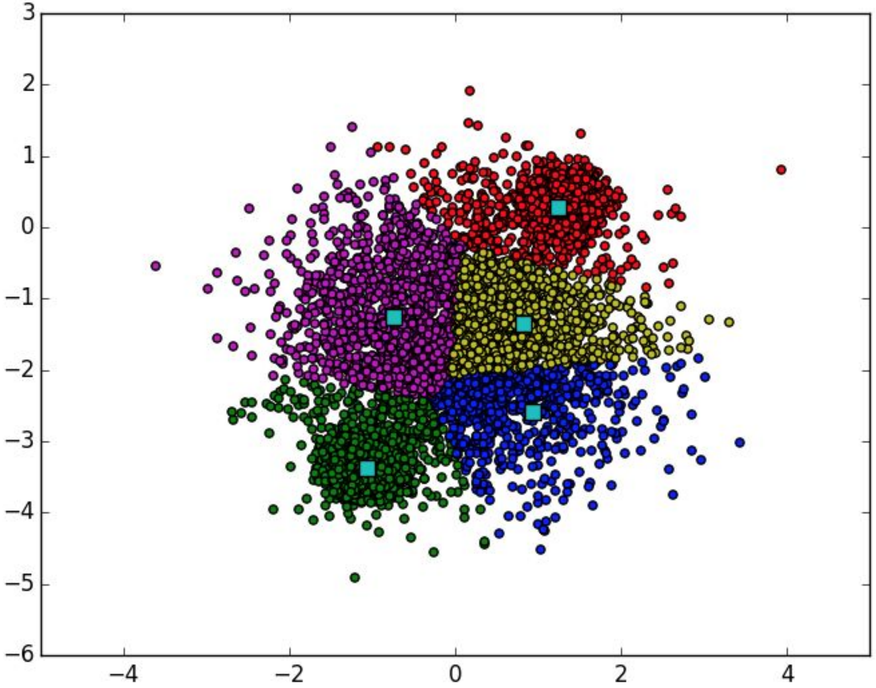
Introduction to Assignment 3



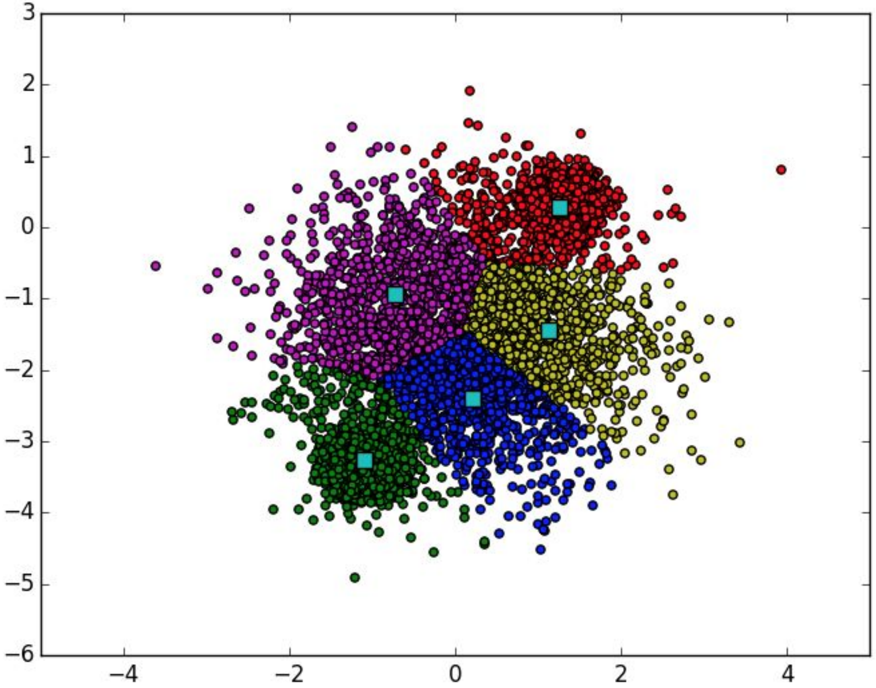
Introduction to Assignment 3



Introduction to Assignment 3



Introduction to Assignment 3



Introduction to Assignment 3

- Gaussian Mixture Model

$$P(\mathbf{X}) = \prod_{n=1}^{B} P(\mathbf{x}_n) = \prod_{n=1}^{B} \sum_{k=1}^{K} P(z_n = k) P(\mathbf{x}_n \mid z_n = k)$$
$$= \prod_{n} \sum_{k} \pi^k \mathcal{N}(\mathbf{x}_n \,;\, \boldsymbol{\mu}^k, {\sigma^k}^2)$$

  - Data: $\mathbf{x}_n$
  - Latent mixture assignment: z_n
  - Gaussian mean and std: $\boldsymbol{\mu}^k, {\sigma^k}^2$

Introduction to Assignment 3

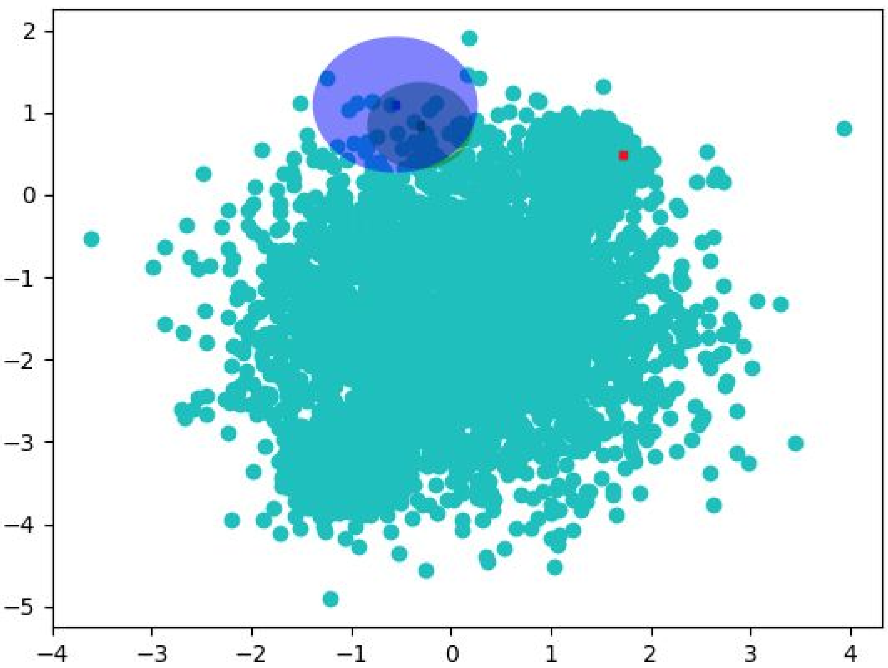
- Numerical Trick:
    - LogSumExp:

$$LSE(x_1, \ldots, x_n) = \log(\exp(x_1) + \cdots + \exp(x_n))$$

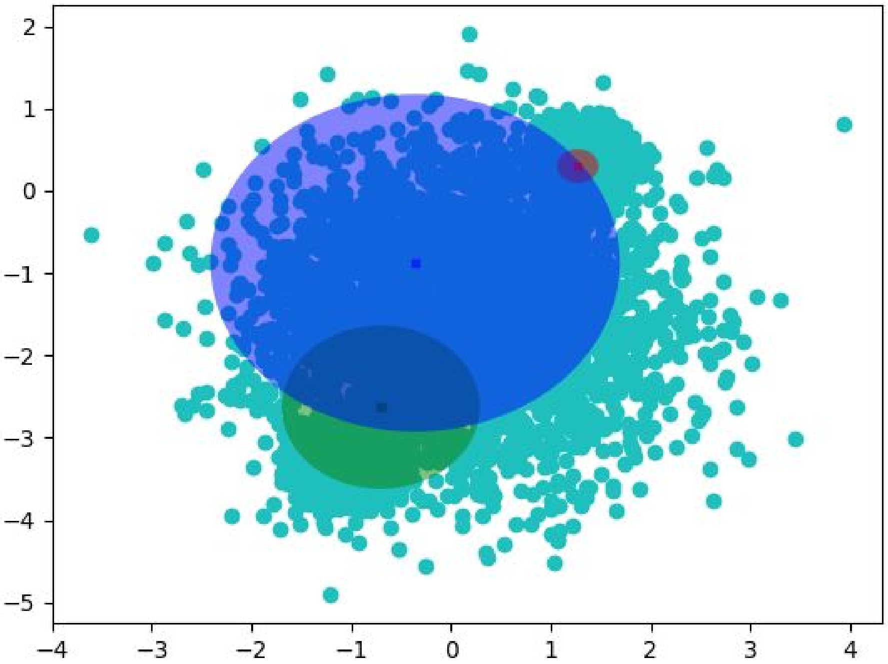
$$LSE(x_1, \ldots, x_n) = x^* + \log(\exp(x_1 - x^*) + \cdots + \exp(x_n - x^*))$$

where $x^* = \max\{x_1, \ldots, x_n\}$

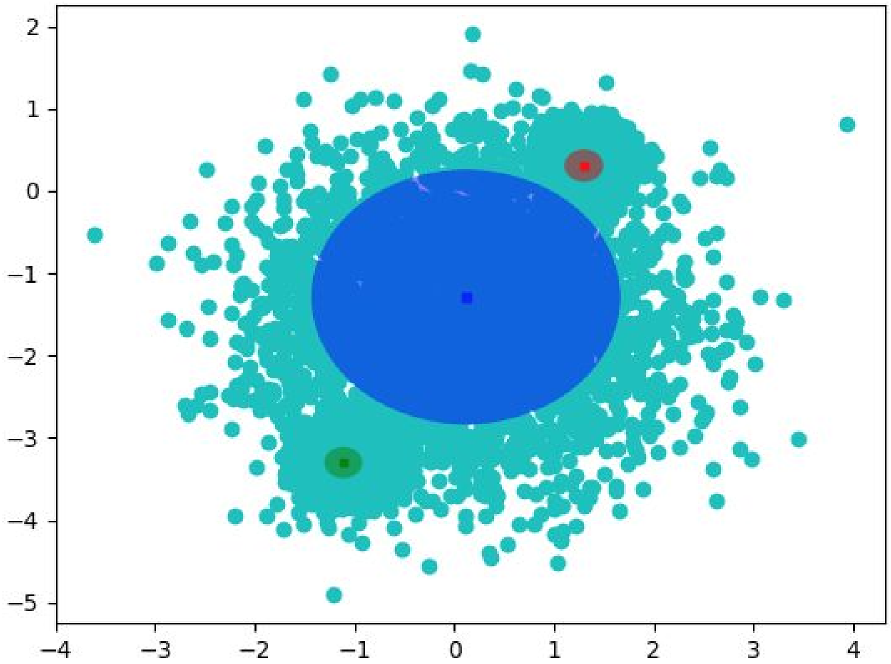
Introduction to Assignment 3



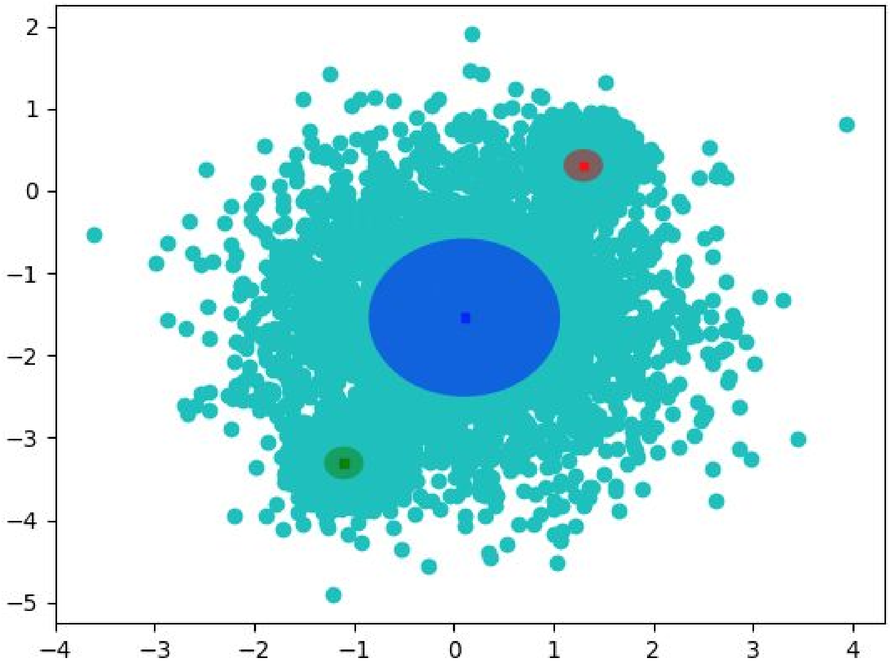
Introduction to Assignment 3



Introduction to Assignment 3



Introduction to Assignment 3



Introduction to Assignment 3

- Factor Analysis

$$P(\mathbf{X}) = \prod_{n=1}^{B} P(\mathbf{x}_n) = \prod_{n=1}^{B} \int_{\mathbf{s}_n} P(\mathbf{s}_n) P(\mathbf{x}_n \,|\, \mathbf{s}_n)$$

$$= \prod_{n=1}^{B} \int_{\mathbf{s}_n} \mathcal{N}(\mathbf{s}_n \,;\, \mathbf{0}, I) \mathcal{N}(\mathbf{x}_n \,;\, W\mathbf{s}_n + \boldsymbol{\mu}, \Psi)$$

$$\log \int_{\mathbf{z}} P(\mathbf{x} \,|\, \mathbf{z}) P(\mathbf{z}) = \mathcal{N}(\mathbf{x} \,;\, \boldsymbol{\mu}, \Psi + WW^T)$$

  - Data: $\mathbf{x}_n$
  - Latent factor: $\mathbf{s}_n$
  - Parameters: W $\boldsymbol{\mu}$, Ψ

Introduction to Assignment 3

- Numerical Trick:
    - Cholesky Decomposition $A = LL^T$

```
log_det = 2.0 * tf.reduce_sum(tf.log(tf.diag_part(tf.cholesky(A))))
```