TU München
Fakultät für Informatik
PD Dr. Rudolph Triebel
John Chiotellis

Machine Learning for Computer Vision
Summer term 2017

August 9, 2017
Topic: HMMs and Sampling methods

Exercise 1: Viterbi algorithm

We play again with our robot from the first homework assignment. As we mentioned back then the robot has a camera with an observation model that looks as follows:

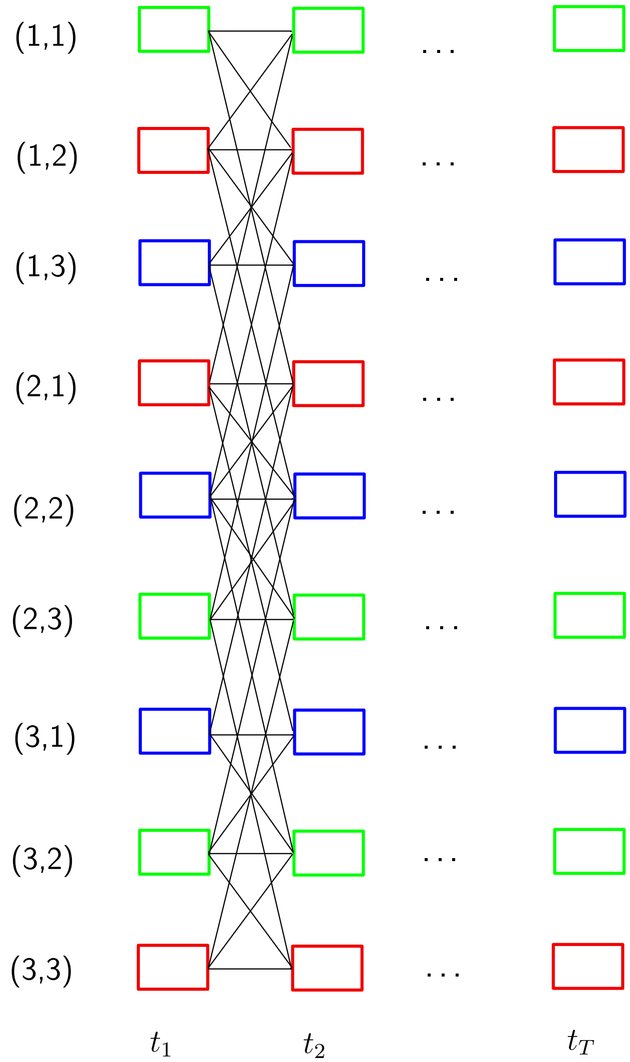
Sensed color \ Actual color	R	G	B
R	0.8	0.1	0.1
G	0.1	0.6	0.2
B	0.1	0.3	0.7

This time we put the robot in a room where the floor looks like this:

y \ x	1	2	3
1	G	R	B
2	R	B	G
3	B	G	R

a) *What is the state space? What is the observation space? Draw the trellis diagram.*

The state is the position of the robot. We have a discrete state space of 9 squares. Each state is a pair (x,y), so $x_i \in \{(1,1),(1,2)\ldots,(3,3)\}$.
The observation space is also discrete and it consists of the 3 colors the robot may observe, so $z_i \in \{R,G,B\}$. The trellis diagram would look as follows:



b) *Assume the robot can only move vertically and horizontally. We let the robot move randomly. If the attempted move leads outside of the bounds of the room the robot stays at its current position. Compute the state transition matrix.*

The robot can only move vertically or horizontally, so there are four possible moves (up, down, left, right). Since the robot moves randomly, each of these has probability $p_{move} = 0.25$. For all states except the one in the central square, there are moves that lead out of the bounds of the room. Then the robot stays at its current position, so the probability for that move is added to the probability of transition to the self-state.

c) *After 3 time steps, what is most likely the path that the robot followed if the camera reads* $\{z_1 =$ *R*, $z_2 =$*G*, $z_3 =$ *G* $\}$ *? Assume the robot's initial position is unknown.*

We want to use the Viterbi algorithm to estimate the most likely sequence of squares the robot followed. To do that we need to compute the transition matrix A (previous question), the initial state probabilities π_i and the observation model $p(z_i|x_i)$. The robot's initial position is unknown, therefore we have $\pi_i = \frac{1}{9} \quad \forall i \in \{1, \ldots, 9\}$.
For the rest, see code.

Exercise 2: Kullback-Leibler divergence

a) *What does the KL divergence describe? What are its key properties?*

The Kullback-Leibler divergence is a measure of (dis-)similarity between probability distributions. It is the extra amount of information needed when a distribution q is used to approximate a distribution p. It is not symmetric and non-negative. It is minimized (zero) when the two distributions are identical. By the definition we have:

$$
\begin{aligned}
KL(p||q) &= \int p(x) \log \frac{p(x)}{q(x)} dx \\
&= \int p(x) \log p(x) dx - \int p(x) \log q(x) dx \\
&= -H(p) + H(p,q) \\
&= \textit{negative entropy of p} + \textit{cross entropy between p and q}
\end{aligned}
$$

b) *Compute the KL-divergence of two univariate normal distributions.*
What if they have the same mean? What if they have the same variance?

Let us define $p_1(x) = \mathcal{N}(x|\mu_1, \sigma_1)$ and $p_2(x) = \mathcal{N}(x|\mu_2, \sigma_2)$. We then have

$$
KL(p_1||p_2) = \int p_1(x) \log\{\frac{p_1(x)}{p_2(x)}\} dx
$$

First let us simplify the fraction

$$
\begin{aligned}
\frac{p_1(x)}{p_2(x)} &= \frac{\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(-\frac{(x-\mu_1)^2}{2\sigma_1^2})}{\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp(-\frac{(x-\mu_2)^2}{2\sigma_2^2})} = \frac{\sigma_2}{\sigma_1} \frac{\exp(-\frac{(x-\mu_1)^2}{2\sigma_1^2})}{\exp(-\frac{(x-\mu_2)^2}{2\sigma_2^2})} \\
&= \frac{\sigma_2}{\sigma_1} \exp(-\frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2})
\end{aligned}
$$

Taking the logarithm of this gives us

$$
\log(\frac{p_1(x)}{p_2(x)}) = \log(\frac{\sigma_2}{\sigma_1}) + \left(\frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2} \right)
$$

Now plugging this in the KL-divergence definition we get

$$
\begin{aligned}
KL(p_1||p_2) &= \int p_1(x)\log(\frac{\sigma_2}{\sigma_1})dx + \int p_1(x)\left(\frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2}\right)dx \\
&= \log(\frac{\sigma_2}{\sigma_1})\int p_1(x)dx + \int p_1(x)\frac{(x-\mu_2)^2}{2\sigma_2^2}dx - \int p_1(x)\frac{(x-\mu_1)^2}{2\sigma_1^2}dx \\
&= \log(\frac{\sigma_2}{\sigma_1}) + \frac{1}{2\sigma_2^2}\int p_1(x)(x-\mu_2)^2dx - \frac{1}{2\sigma_1^2}\int p_1(x)(x-\mu_1)^2dx \\
&= \log(\frac{\sigma_2}{\sigma_1}) + \frac{1}{2\sigma_2^2}\int p_1(x)(x-\mu_1+\mu_1-\mu_2)^2dx - \frac{\sigma_1^2}{2\sigma_1^2} \\
&= \log(\frac{\sigma_2}{\sigma_1}) + \frac{1}{2\sigma_2^2}\left(\int p_1(x)(x-\mu_1)^2dx + 2\int p_1(x)(x-\mu_1)(\mu_1-\mu_2)dx + \int p_1(x)(\mu_1-\mu_2)^2dx\right) - \frac{1}{2} \\
&= \log(\frac{\sigma_2}{\sigma_1}) + \frac{1}{2\sigma_2^2}\left(\sigma_1^2 + 2(\mu_1-\mu_2)\int p_1(x)(x-\mu_1)dx + (\mu_1-\mu_2)^2\int p_1(x)dx\right) - \frac{1}{2} \\
&= \log(\frac{\sigma_2}{\sigma_1}) + \frac{1}{2\sigma_2^2}\left(\sigma_1^2 + (\mu_1-\mu_2)^2\right) - \frac{1}{2}
\end{aligned}
$$

If two distributions only differ in their mean values ($\sigma_1 = \sigma_2$) then the KL-divergence is proportional to the square of their means difference,

$$KL(p||q) = \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2}.$$

If they have equal mean but different variances ($\mu_1 = \mu_2$) then the KL-divergence is a function of the ratio of their variances:

$$KL(p||q) = \log(\frac{\sigma_2}{\sigma_1}) + \frac{\sigma_1^2}{2\sigma_2^2} - \frac{1}{2} = \frac{\sigma_1^2}{2\sigma_2^2} - \log(\frac{\sigma_1}{\sigma_2}) - \frac{1}{2}$$

c) *Consider a factorized variational distribution $q(Z)$. By using the technique of Lagrange multipliers, verify that minimization of $KL(p||q)$ with respect to one of the factors $q_i(Z_i)$ keeping all other factors fixed, leads to the solution:*

$$q_j^*(Z_j) = \int p(Z)\prod_{i\neq j} dZ_i = p(Z_j)$$

$$
\begin{aligned}
KL(p||q) &= \int p(Z) \ln \frac{p(Z)}{q(Z)} dZ \\
&= \int p(Z) \ln p(Z) dZ - \int p(Z) \ln q(Z) dZ \\
&= \int p(Z) \ln p(Z) dZ - \int p(Z) \ln \prod_i q_i(Z_i) dZ \\
&= -\int p(Z) \sum_{i=1}^{M} \ln q_i(Z_i) dZ + const. \\
&= -\int (p(Z) \ln q_j(Z_j) + p(Z) \sum_{i \neq j} \ln q_i(Z_i)) dZ + const. \\
&= -\int p(Z) \ln q_j(Z_j) dZ + const. \\
&= -\int \ln q_j(Z_j) \left(\int p(Z) \prod_{i \neq j} dZ_i \right) dZ_j + const.
\end{aligned}
$$

Note that by *const.* we imply w.r.t. q_j. We want to minimize this and at the same time enforce the constraint

$$
\int q_j(Z_j) dZ_j = 1.
$$

Therefore we add a Lagrange multiplier and our objective function becomes

$$
\mathcal{L}(q_j(Z_j)) = -\int \ln q_j(Z_j) \left(\int p(Z) \prod_{i \neq j} dZ_i \right) dZ_j + \lambda \left(\int q_j(Z_j) dZ_j - 1 \right)
$$

Taking the derivative w.r.t. $q_j(Z_j)$ and setting it equal to zero we get

$$
\frac{\partial \mathcal{L}(q_j(Z_j))}{\partial q_j(Z_j)} = -\frac{\int p(Z) \prod_{i \neq j} dZ_i}{q_j(Z_j)} + \lambda \overset{!}{=} 0
$$

We solve for λ

$$
\begin{aligned}
\lambda q_j(Z_j) &= \int p(Z) \prod_{i \neq j} dZ_i \\
\lambda \int q_j(Z_j) dZ_j &= \int \left(\int p(Z) \prod_{i \neq j} dZ_i \right) dZ_j \\
\lambda &= 1
\end{aligned}
$$

And thus

$$
q_j^*(Z_j) = \int p(Z) \prod_{i \neq j} dZ_i = p(Z_j)
$$

Exercise 3: Particle Filter

a) *What kind of spaces can we explore with a particle filter?*

With particle filters we can explore continuous state spaces.

b) *What kind of distributions can we approximate with a particle filter?*

Particle filter is non-parametric, meaning we can approximate arbitrary distributions (Gaussian and non-Gaussian). Given enough particles we can approximate any function.

c) *In a Monte Carlo localization problem what do the particles and the particle weights correspond to?*

The particles themselves correspond to the motion model as they represent the state after motion with noise. The particle weights are computed according to the measurement model so they represent the likelihood of a measurement.

d) **Programming** : *Implement a particle filter for global localization.*
See code.

Exercise 4: Gibbs sampling

Show that the Gibbs sampling algorithm satisfies detailed balance:

$$p(z)T(z, z') = p(z')T(z', z)$$

This follows from the fact that in Gibbs sampling, we sample a single variable, z_k at each time, while all other variables, $z_{-k} = \{z_i\}_{i \neq k}$, remain unchanged. Thus, $z'_{-k} = z_{-k}$. We denote as $T(z, z')$ the transition probability from z to z' and we get

$$\begin{aligned}
p(z)T(z, z') &= p(z_k, z_{-k})p(z'_k|z_{-k}) && \text{(Joint probability)} \\
&= p(z_k|z_{-k})p(z_{-k})p(z'_k|z_{-k}) && \text{(Product Rule)} \\
&= p(z_k|z'_{-k})p(z'_{-k})p(z'_k|z'_{-k}) && (z_{-k} = z'_{-k}) \\
&= p(z_k|z'_{-k})p(z'_k, z'_{-k}) && \text{(Product Rule)} \\
&= T(z', z)p(z') && \text{(Joint probability)},
\end{aligned}$$

where we have used the transition probability in Gibbs sampling $T(z, z') = p(z'_k|z_{-k})$.