Machine Learning WS2014/15 (Unsupervised Learning)

Lecturer: Prof. Dr. Bethge

1. Expectations, information theory and Gaussians

- **Joint distribution:**

$$p(x_i, y_j) = P(\{s \in S : X(s) = x_i\} \cap \{s \in S : Y(s) = y_j\})$$

The following properties hold:

(i) $\sum_i p(x_i, y_j) = p(y_j)$

(ii) $\sum_j p(x_i, y_j) = p(x_i)$

(iii) $\sum_i \sum_j p(x_i, y_j) = 1$

- **Joint cumulative distribution function:**

$$F(x) = P(\{s \in S : X(s) \leq x\} \cap \{s \in S : Y(s) \leq y\}), \quad \forall (x, y) \in (S_x \times S_y)$$

The following properties hold:

(i) F is a nondecreasing function in x and y.

(ii) $\lim_{x \to -\infty} \lim_{y \to -\infty} F(x, y) = 0, \quad \lim_{x \to \infty} \lim_{y \to \infty} F(x, y) = 1$

(iii) $P(\{s \in S : X(s) > x\} \cap \{s \in S : Y(s) \leq y\})) = F(\infty, y) - F(x, y)$,
$P(\{s \in S : X(s) > x\} \cap \{s \in S : Y(s) > y\})) = 1 - F(\infty, y) - F(x, \infty) + F(x, y)$

(iv) $P(\{s \in S : x_1 < X(s) \leq x_2\} \cap \{s \in S : y_1 < Y(s) \leq y_2\}) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1)$

- **Expectation:** $E[f(X)] := \begin{cases} \sum p(x) f(x) & , \text{ if p is a probability mass function} \\ \int p(x) f(x) dx & , \text{ if p is a density function} \end{cases}$

- **Covariance:** $Cov[X, Y] = E[XY] - E[X]E[Y]$,

- **Covariance matrix:** for n random variables $X_1, \ldots, X_n$ the *Covariance matrix* is defined by: $C_{ij} := Cov[X_i, X_j] = E[X_i X_j] - E[X_i]E[X_j]$

- **Independence:** Two random variables X, Y are *(statistically) independent* if their joint distribution is factorial: $p(x, y) = p(x)p(y)$ or, equivalently, if their joint cdf is factorial: $F(x, y) = F(x)F(y)$

- **Kullback-Leibler divergence**

$$D_{KL}[p(z)||\hat{p}(z)] = \begin{cases} \sum_z p(z) \log \frac{p(z)}{\hat{p}(z)} & , \text{ if p is a probability mass function} \\ \int p(z) \log \frac{p(z)}{\hat{p}(z)} dz & , \text{ if p is a density function} \end{cases}$$

- **Mutual information**

$$I[X : Y] := D_{KL}[p(x, y)||p(x)p(y)] = \begin{cases} \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} & , \text{ if p is a probability mass function} \\ \int p(x, y) \log \frac{p(x,y)}{p(x)p(y)} dxdy & , \text{ if p is a density function} \end{cases}$$

Task 1:
Show the following properties:

1) The expectation value is a linear form: $E[aX + bY + c] = aE[X] + bE[Y] + c$
2) $Var[X] = Cov[X, X]$
3) $Cov[X, Y] = E[(X - E[X])(Y - E[Y])]$ and hence $Var[X] = E[(X - E[X])^2]$
4) The covariance is a bilinear form $Cov[aX + b, cY + d] = ac\,Cov[X, Y]$ and hence $Var[aX] = a^2\,Var[X]$
5) If X, Y independent, then $Cov[X, Y] = 0$
6) The inversion is not warranted. Counter example: $p(x, y) = 1/8$ for all integers x, y for which $|x| + |y| = 2$ and zero otherwise. Then $Cov[X, Y] = 0$ but $p(x, y) \neq p(x)p(y)$.
7) $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$ or more generally:
$Var\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n}\sum_{j=1}^{n} Cov[X_k, X_j] = \sum_{k=1}^{n} Var[X_k] + \sum_{j\neq k} Cov[X_k, X_j]$
8) $(Cov[X, Y])^2 \leq Var[X]\,Var[Y]$
(Hint: Let $\tilde{X} := X - E[X]$, $\tilde{Y} := Y - E[Y]$, and $Z := (\tilde{X} - t\tilde{Y})^2$. Then $E[Z] = t^2E[\tilde{Y}^2] - 2tE[\tilde{X}\tilde{Y}] + E[\tilde{X}^2] \geq 0\forall t$ and hence $E[\tilde{X}\tilde{Y}]^2 \leq E[\tilde{X}^2]E[\tilde{Y}^2]$ because $at^2 + bt + c \geq 0\ \forall t \ \Leftrightarrow\ b^2 \leq 4ac$.)
9) Decomposition of *total variance*: $Var[X] = E[Var[X|Y]] + Var[E[X|Y]]$

Task 2 (information theory):
For multivariate random variables $X \in R^m$ and $Y \in R^n$ show the following properties:

1) $I[X : Y] = h[X] + h[Y] - h[X, Y]$
2) $I[X : Y] = h[X] - h[X|Y]$
3) $h[X, Y] = h[X] + h[Y|X]$
4) $h[Y] = h[X] + E\left[\log\left|\left(\frac{\partial y_j}{\partial x_k}\right)\right|\right]$ where $Y = f(X)$ and $\left(\frac{\partial y_j}{\partial x_k}\right)$ denotes the Jacobian of $f(x)$.
5) $h[\mathcal{N}(\mu, \sigma^2)] \equiv E[-\log\mathcal{N}(x|\mu, \sigma^2)] \equiv -\int\mathcal{N}(x|\mu, \sigma^2)\log\mathcal{N}(x|\mu, \sigma^2)dx = \frac{1}{2}\log 2\pi e\sigma^2$
6) $h[\mathcal{N}(\mu, C)] = \frac{1}{2}\log(2\pi e)^D|C|$ where $C \in R^{D\times D}$ and $|C|$ denotes the (absolute value of the) determinant of C.
7) For $\mathbf{y} \sim \mathcal{N}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}D_{11} & 0\\ 0 & D_{22}\end{pmatrix}\right)$, $r(\phi) := \begin{pmatrix}\cos\phi\\ \sin\phi\end{pmatrix}$ and $z_\phi := r^\top x$ compute and plot the following two functions: $f(\phi) := Var[z_\phi]$ and $g(\phi) := I[z_\phi : y]$.

Task 3:
Find errors in ML_script_1.pdf