Module: M2. Optimization and inference techniques for Computer Vision **Final exam**
Date: December 1rst, 2016
Teachers: Juan Fco Garamendi, Coloma Ballester, Oriol Ramos, Joan Serrat **Time: 2h30min**

- Books, lecture notes, calculators, phones, etc. are not allowed.
- All sheets of paper should have your name.
- Answer each problem in a separate sheet of paper.
- All results should be demonstrated or justified.

Problem 1

Juan F. Garamendi, 1 Point

Let

$$\begin{aligned} J: \quad \mathcal{V} &\to \mathbb{R}, \\ u &\mapsto J(u) = \int_\Omega \mathcal{F}\left(\mathbf{x}, u(\mathbf{x}), \nabla u(\mathbf{x})\right) d\mathbf{x} \end{aligned}$$

be a convex energy functional over functions u, where

- $\mathcal{V}$ is a suitable space of functions.
- $\Omega \in \mathbb{R}^d$ is a bounded open domain of the d dimensional euclidean space $\mathbb{R}^d$.
- $u \in \mathcal{V}$, $u : \Omega \to \mathbb{R}$ is a scalar function defined on Ω.
- $\mathbf{x} \in \Omega$ such that $\mathbf{x} = (x_1, .., x_d)$ is the spatial variable and ∇ is the gradient operator such that $\nabla u(\mathbf{x}) = (u_{x_1}, .., u_{x_d})$

(a) (0.25 points) Say in few words which is the fundamental problem in calculus of variations.

The fundamental problem of the calculus of variations is to find the extremum (maximum or minimum) of the functional $J(u)$ with respect to u.

(b) (0.25 points) Write the definition of the Gâteaux derivative of $J(u)$ (the directional derivative of J at u in direction of function $h(\mathbf{x})$

$$\left.\frac{d\,J}{d\,u}\right|_h = \lim_{\alpha \to 0} \frac{J(u + \alpha h) - J(u)}{\alpha}$$

(c) (0.25 points) Let

$$\frac{d\,J}{d\,u}$$

be the derivative of J at u. Which is the necessary condition for extremality of J?

$$\frac{d\,J}{d\,u} = 0$$

(d) (0.25 points) Explain in a few words the relationship between the Gâteaux derivative and the following expressions:

(i)
$$-\text{div}_{\mathbf{x}}\left(\nabla_{\bar{g}}\mathcal{F}\right) + \frac{\partial \mathcal{F}}{\partial u} = 0$$

where

- The divergence $\text{div}_{\mathbf{x}}$ is computed with respect to variable $\mathbf{x}$
- The gradient $\nabla_{\bar{g}}$ is computed with respect to the components of $\bar{g} = \nabla u(\mathbf{x})$

(ii)
$$-\sum_{i=1}^{d} \frac{\partial}{\partial x_i}\frac{\partial \mathcal{F}}{\partial u_{x_i}} + \frac{\partial \mathcal{F}}{\partial u} = 0$$

The both expressions are the canonical (general) expression for the Euler-Lagrange equation and they are the necessary condition for the extremality of J applying the definition of the Gâteaux derivative to J.

Problem 2

Juan F. Garamendi, 1 Points

For binary segmentation, the model problem can be defined as follows: Let $\omega \subset \Omega$ be an open, positive measured sub-region of the original domain (eventually not connected). If the curve Γ represents the boundary of such a segmentation ω then, in the level set formulation, the (free) moving boundary Γ is the zero level set of a Lipschitz function $\phi : \Omega \to \mathbb{R}$, that is:

$$\Gamma = \{(x, y) \in \Omega : \phi(x, y) = 0\}$$

where

$$\omega = \{(x, y) \in \Omega : \phi(x, y) > 0\}$$

and

$$\Omega \backslash \overline{\omega} = \{(x, y) \in \Omega : \phi(x, y) < 0\}$$

The level set function ϕ can be characterized as a minimum of the following energy functional,

$$J(\bar{c}, \phi) = \int_{\Omega} |DH(\phi)| + \frac{1}{2\lambda}\int_{\Omega} (f - c_1)^2 H(\phi) \ + (f - c_2)^2 (1 - H(\phi)) d\mathbf{x}$$

where $DH(\phi)$ is the distributional gradient of Heaviside function $H(\phi)$, f is a bounded function representing the image (the data), $\bar{c} \in \mathbb{R}^2$, $\bar{c} = (c_1, c_2)$ are a vectorial unknown values representing the classes, and $\lambda \in \mathbb{R}^+$ is a given parameter. The function $H(x)$ represents the Heaviside function, i.e.: $H(x) = 1$ if $x \geq 0$ and $H(x) = 0$ otherwise, and it allows to express the length of Γ by

$$|\Gamma| = \text{Length}(\phi = 0) = \int_{\Omega} |DH(\phi)|$$

where the term $\int_{\Omega} |DH(\phi)|$ denotes, properly, the total variation of the discontinuous function $H(\phi)$ in Ω.

(a) (0.15 points) Why c_1 and c_2 are the mean values of the inside and outside regions?

Because when the energy functional is minimized w.r.t. c_1 and c_2, the necessary condition that is derivative of the energy functional w.r.t. c_1 and c_2 is zero give us

$$c_1 = \frac{\int_{\Omega} f H(\phi) d\mathbf{x}}{\int_{\Omega} H(\phi) d\mathbf{x}} \quad c_2 = \frac{\int_{\Omega} f\,(1 - H(\phi))\, d\mathbf{x}}{\int_{\Omega} 1 - H(\phi) d\mathbf{x}}$$

That are the the mean values.

(b) (0.85 points) Let $c_1, c_2 \in \mathbb{R}$, $c_1 > c_2$, be known values, $\lambda \in \mathbb{R}^+$ be a positive given parameter and $u \in BV(\Omega$ be the (unique) minimum of the Rudin-Osher-Fatemi energy functional

$$J_{rof}(u) = \int_\Omega |Du| + \frac{1}{2}\lambda_{rof} \int_\Omega |u - f|^2 d\mathbf{x}$$

where $\lambda_{rof} = \frac{c_1 - c_2}{\lambda}$.

Chambolle and Darbon proved[1] that the set $\omega = \{\mathbf{x} \in \Omega | u(\mathbf{x}) > (c_1 + c_2)/2\}$ is a solution of

$$J(\phi) = \int_\Omega |DH(\phi)| + \frac{1}{2\lambda} \int_\Omega (f - c_1)^2 H(\phi) \ + (f - c_2)^2 (1 - H(\phi)) d\mathbf{x}$$

with $\Gamma = \partial\, \omega = \{\mathbf{x} \in \Omega | \phi(\mathbf{x}) = 0\}$.

Use this information to complete the following algorithm to minimize energy functional $J(\bar{c}, \phi)$ minimizing the Rudin-Osher-Fatemi $J_{rof}(u)$ energy.

(i) Fix an initial partition ω.

(ii) Minimize the energy $J(\bar{c}, \phi)$ with respect to the constant variables c_1, c_2, i.e. choose c_1 and c_2 as the mean values inside ω and outside ω respectively. Check if $c_1 > c_2$, if it not the case, just exchange names.

(iii)

$$\vdots$$

(iv)

$$\vdots$$

(i) Fix an initial partition ω.

(ii) Minimize the energy $J(\bar{c}, \phi)$ with respect to the constant variables c_1, c_2, i.e. choose c_1 and c_2 as the mean values inside ω and outside ω respectively. Check if $c_1 > c_2$, if it not the case, just exchange names.

(iii) Minimize

$$J_{rof}(u) = \int_\Omega |Du| + \frac{1}{2}\lambda_{rof} \int_\Omega |u - f|^2 d\mathbf{x}$$

with $\lambda_{rof} = \frac{c_1 - c_2}{\lambda}$. Then, the set $\omega = \{\mathbf{x} \in \Omega | u(\mathbf{x}) > (c_1 + c_2)/2\}$, minimize $J(\bar{c}, \phi)$ w.r.t. ϕ in such a way $\omega = \{\mathbf{x} \in \Omega | \phi(\mathbf{x}) > 0\}$,

(iv) Go to step 2 until convergence

Problem 3

Juan F. Garamendi, 0.5 Points

Consider the following iterative scheme

- $\bar{x}^k \leftarrow S_i(\bar{x}^{k-1}, \bar{x}^k, \bar{b})$

used to solve the algebraic problem $\mathbf{A}\bar{x} = \bar{b}$, i.e., $\lim_{k\to\infty} \bar{x}^k = \bar{x}$, where $\mathbf{A}$ is a known matrix, $\bar{b}$ is a known vector, $\bar{x}$ is an unknown vector , S_i some function , super-index represents iteration number and $\bar{x}^0$ some initial value. Now consider two versions of S_i: S_1, S_2 with the following behaviour

(a) $e = \bar{x} - S_1(\bar{x}^1, \bar{x}^2, \bar{b})$, $\|e\|_\infty = 10^{-1}$, with e having only high frequencies.

(b) $e = \bar{x} - S_2(\bar{x}^1, \bar{x}^2, \bar{b})$, $\|e\|_\infty = 10^{1}$, with e having only low frequencies.

[1] A. Chambolle and J. Darbon, *On Total Variation Minimization and surface evolution using parametric maximum flows*. Int. Journal of Computer Vision. 84,3 (2009) pp 288-307

(a) (0.5 points) Explain in few words which is the best iterative scheme (S_1 or S_2) for embedding it into a multigrid scheme.

In a multigrid scheme the iterative scheme S, used to obtain the approximated solution $\bar{x}^k$, must have a remarkable smoothing effect on the error $e = \bar{x} - \bar{x}^k$. This is because the residual equation $\mathbf{A}\bar{e} = \bar{r}$, being $\bar{r} = \bar{b} - \mathbf{A}\bar{x}^k$ the residual, will be solved in a coarser level than the original algebraic problem $\mathbf{A}\bar{x} = \bar{b}$, so it is important that $\bar{e}$ in the coarser level represents well the error $\bar{e}$ at the finer level, and this is only possible if $\bar{e}$ does not have high frequencies (We want to avoid the aliasing effect), i.e. the best scheme is S_2.

Problem 4

Coloma Ballester *0.75 Points*

The following Figure 1 shows 500 measures, $(t_1, y_1), \ldots, (t_{500}, y_{500})$ (where $t_i \in \mathbb{R}$, $y_i \in \mathbb{R}$, $i = 1, \ldots, 500$), which have been taken during the first trimester of 2016, related to the value of stock market shares of a company (a Computer Vision company).

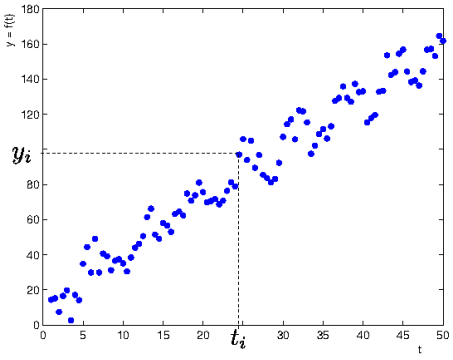


Figure 1: The data.

We assume that there is a functional dependence between t and y, $y = f(t)$, and thus we would like to fit a function f to this given data set. We also assume that f can be modelled as

$$f(t) = at + b + c\cos t,$$

which depends on the (unknown) parameters $a, b, c \in \mathbb{R}$.

(a) Explain the least squares solution to this problem of data fitting, used to determine the parameters $\mathbf{x}^t = [a, b, c]^t$. Write down the expression of the energy $E(\mathbf{x})$ (or $E(a, b, c)$) to be minimized.

When fitting the function $f(t) = at + b + c\cos t$,, to the data $(t_1, y_1), \ldots, (t_N, y_N)$, the coeffients a, b, c are unknowns we have to determine. Notice that we talk about linear regression because, although f is not a linear function of t, it is indeed linear on the coeffients a, b, c.

There are more data than unknowns, $500 > 3$. We compute the *best* unknowns a, b, c by minimizing the sum of squared errors, between the predicted value $f(t)$ and the measured value, y, over the data:

$$\min_{a,b,c} \sum_{i=1}^{500} |f(t_i) - y_i|^2 = \min_{a,b,c} \sum_{i=1}^{500} |at_i + b + c\cos t_i - y_i|^2$$

Therefore, the energy to be minimized is $E(a, b, c) = \sum_{i=1}^{500} |at_i + b + c\cos t_i - y_i|^2$. We can express E, and the above problem, in matrix notation as

$$\min_{\mathbf{x}} E(\mathbf{x}) = \min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{y}\|^2 \tag{1}$$

where $\mathbf{y} = [y_1, y_2, \ldots, y_{500}]^t \in R^{500}$, $\mathbf{x} = [a, b, c]^t \in \mathbb{R}^3$ and

$$A = \begin{bmatrix} t_1 & 1 & \cos t_1 \\ t_2 & 1 & \cos t_2 \\ \vdots & \vdots & \vdots \\ t_{500} & 1 & \cos t_{500} \end{bmatrix}$$

is an 500×3 matrix (called the *design matrix*).

(b) Write down the normal equations associated to this problem.

The normal equations associated to the least squares problem (1) are $A^t A\mathbf{x} = A^t\mathbf{y}$.

(c) How can you solve the normal equations and determine the parameters $\mathbf{x}^t = [a, b, c]^t$ using the SVD or the pseudoinverse of the matrix associated to your problem?
(Recall that SVD stands for Singular Value Decomposition of a matrix).

The solution to the normal equations with minimum Euclidian norm is $\mathbf{x}_0 = A^+\mathbf{y}$, where A^+ is the pseudoinverse of A.
To compute the pseudoinverse, we use the SVD decomposition of A, which in turn is:

$$A = U\Sigma V^t$$

where U is a 500×500 orthogonal matrix, V is a 3×3 orthogonal matrix, and Σ is a rectangular 500×3 diagonal matrix. The non-zero diagonal values of Σ, $\sigma_1, \ldots, \sigma_r$, are called the singular values and they are all strictly positive ($\sigma_i > 0$). The number of singular values r is the rank of A (thus, $r \leqslant 3 = \min\{500, 3\}$).
From the SVD of A, we compute its pseudoinverse:

$$A^+ = V(\Sigma^t)^+ U^t$$

where the pseudo-inverse $(\Sigma^t)^+$ is a 3×500 diagonal matrix with entries

$$\sigma_i^+ = \begin{cases} \frac{1}{s_{ii}} & \text{if } s_{ii} (= \sigma_i) \neq 0 \\ 0 & \text{if } sii = 0, \end{cases}$$

(where s_{ii} are the diagonal entries of Σ), for $i = 1, 2, 3$.

Problem 5 Coloma Ballester *1. Points*

Consider the function $f : \mathbb{R}^n \to \mathbb{R}$, $n \in \mathbb{N}$, defined by

$$f(x) = \frac{1}{2}\langle x, Px\rangle - \langle x, q\rangle,$$

for $x \in \mathbb{R}^n$, where P is a $n \times n$ symmetric matrix and $q \in \mathbb{R}^n$. (The notation $\langle \cdot, \cdot \rangle$ stands for the Euclidian scalar product in $\mathbb{R}^n$.)

(a) What condition on the matrix P implies that f is a convex function?

The answer is P positive definite (that is, $\langle Px, x\rangle \geq 0$ for all x or, equivalently, all eigenvalues of P are positive).
Indeed, we know that, for a differentiable function, f is convex if and only if all eigenvalues of its Hessian $D^2 f(x)$ are positive for all x. In our case, $D^2 f(x_0) = P$.

(b) Let $v \in R^n, v \neq 0$. Compute $D_v f(x)$, the directional derivative of f at the point x in the direction v. Use this result to compute $\nabla f(x)$.

Given $v \in R^n, v \neq 0$, and $x \in R^n$:

$$D_v f(x) = \lim_{\epsilon \to 0+} \frac{f(x+\epsilon v) - f(x)}{\epsilon} = \lim_{\epsilon \to 0+} \frac{\frac{1}{2}\langle x+\epsilon v, P(x+\epsilon v)\rangle - \langle x+\epsilon v, q\rangle - \left(\frac{1}{2}\langle x, Px\rangle - \langle x, q\rangle\right)}{\epsilon} =$$

$$= \lim_{\epsilon \to 0+} \frac{\frac{\epsilon}{2}\langle x, Pv\rangle + \frac{\epsilon}{2}\langle v, Px\rangle + \frac{\epsilon^2}{2}\langle v, Pv\rangle - \epsilon\langle v, q\rangle}{\epsilon} = \frac{1}{2}\langle P^t x, v\rangle + \frac{1}{2}\langle Px, v\rangle - \langle q, v\rangle = \langle Px - q, v\rangle.$$

We know that

$$D_v f(x_0) = \langle \nabla f(x), v\rangle.$$

for all $v \in R^n$. We can extract from this the vector $\nabla f(x)$:

$$\nabla f(x) = Px - q.$$

(c) Assuming that you are in the situation where f is convex and that P is invertible, verify that the value $x_0 \in \mathbb{R}^n$ where the minimum is attained is $x_0 = P^{-1}q$. Justify the argument you are using.

The necessary condition for a minimum of $f(x)$ is $\nabla f(x) = 0$ (Euler-Lagrange equation). In our case, $Px = q$.

As we are assuming that f is convex (for instance because P is positive definite, as in (a)), in $\mathbb{R}^n$ it is a sufficient condition.

Then, as P is invertible, the minimum is $x_0 = P^{-1}q$.

Problem 6

Coloma Ballester *0.75 Points*

Let A be a $m \times n$ matrix, and $b \in \mathbb{R}^n$, $m, n \in \mathbb{N}$. Consider the problem of minimizing the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by

$$f(x) = \|Ax\|_{\mathbb{R}^m} + \frac{1}{2\lambda}\|x - b\|^2_{\mathbb{R}^n}$$

for $\lambda > 0$. The notation $\|\cdot\|_{\mathbb{R}^k}$ stands for the Euclidean norm in $\mathbb{R}^k$, for $k \in \mathbb{N}$. Note that f is not differentiable when $Ax = 0$.

(a) Write the problem

$$\min_x f(x) \qquad \text{(P)}$$

as a min-max problem. Define also the duality gap.
Hint: Remember the fact that $\|y\|_{\mathbb{R}^k} = \max_{\|\xi\|_{\mathbb{R}^k} \leq 1} \langle y, \xi\rangle_{\mathbb{R}^k}$.

Using that $\|Ax\|_{\mathbb{R}^m} = \max_{\xi \in C} \langle Ax, \xi\rangle_{\mathbb{R}^m}$, where $C = \{\xi \in \mathbb{R}^m : \|\xi\|_{\mathbb{R}^m} \leq 1\}$, we have that:

$$f(x) = \max_{\xi \in C} \left(\langle Ax, \xi\rangle_{\mathbb{R}^m} + \frac{1}{2\lambda}\|x - b\|^2_{\mathbb{R}^n} \right),$$

Then:

$$\min_x f(x) = \min_x \max_{\xi \in C} \left(\langle Ax, \xi\rangle_{\mathbb{R}^m} + \frac{1}{2\lambda}\|x - b\|^2_{\mathbb{R}^n} \right),$$

The duality gap is the difference

$$DG = \min_{x \in R^n} \max_{\xi \in C} \left(\langle Ax, \xi\rangle + \frac{1}{2\lambda}\|x - b\| \right) - \max_{\xi \in C} \min_{x \in R^n} \left(\langle Ax, \xi\rangle + \frac{1}{2\lambda}\|x - b\| \right).$$

(b) What is the primal-dual problem for problem (P)? Are they equivalent problems?

As the function

$$\mathcal{L}(x,\xi) = \langle Ax, \xi\rangle + \frac{1}{2\lambda}\|x-b\|,$$

depending on the primal variables x and on the dual variables ξ, is convex with respect x (for each $\xi \in C$ fixed) and concave with respect to ξ (for each $x \in \mathbb{R}^n$ fixed), then $DG = 0$ and the three problems (the Primal problem (P), the Dual problem, and the Primal-Dual problem) are equivalent.

The Primal-Dual problem is

$$\min_{x\in R^n}\max_{\xi\in C}\left(\langle Ax,\xi\rangle + \frac{1}{2\lambda}\|x-b\|\right) = \max_{\xi\in C}\min_{x\in R^n}\left(\langle Ax,\xi\rangle + \frac{1}{2\lambda}\|x-b\|\right)$$

(c) Define and compute the dual function and the dual problem of problem (P).

The dual function is $g_D(\xi) = \mathcal{L}(x_0(\xi),\xi)$, where

$$x_0(\xi) = \arg\min_x \mathcal{L}(x,\xi) = \arg\min_x \left(\langle Ax,\xi\rangle + \frac{1}{2\lambda}\|x-b\|\right).$$

The minimizer $x_0(\xi)$ is the solution of $\nabla_x \mathcal{L}(x,\xi) = 0$, which is

$$x_0(\xi) = b - \lambda A^t \xi.$$

Substituting $x_0(\xi)$ one obtains the dual function:

$$g_D(\xi) = \mathcal{L}(x_0(\xi),\xi) = \langle Ab,\xi\rangle_{\mathbb{R}^m} - \frac{\lambda}{2}\|A^t\xi\|^2_{\mathbb{R}^m}.$$

Finally the dual problem is

$$\max_{\xi\in C} g_D(\xi) = \max_{\xi\in C}\langle Ab,\xi\rangle_{\mathbb{R}^m} - \frac{\lambda}{2}\|A^t\xi\|^2_{\mathbb{R}^m}.$$

(which is a quadratic problem with constraints, where we have eliminated the primal variable, and therefore could be solved with a projected gradient ascent).

Problem 7 Joan Serrat *1 Point*

We saw that binary image denoising could be modeled as a problem of maximum a posteriori inference over the graphical model of Figure 2 below. The goal then was

$$\arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) = \arg\max_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})\; p(\mathbf{x})$$

(a) What are $\mathbf{x}$ and $\mathbf{y}$?

(b) What are the names for the $p(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{x})$ terms ?

(c) Which is the "order" of the model ? (can be a number or a word)

(d) The term $p(\mathbf{y})$ does not appear, how comes we can get rid of it ?

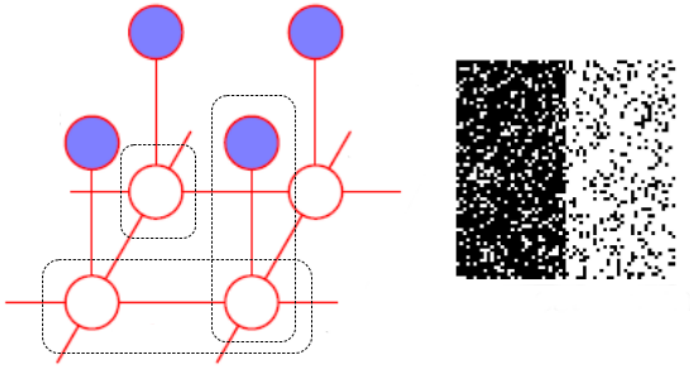


Figure 2: Graphical model for binary image denoising.

(a) $\mathbf{x}$ is the sought clean image, $\mathbf{y}$ is the observation or noisy image we have

(b) $p(\mathbf{y}|\mathbf{x})$ is the likelihood, $p(\mathbf{x})$ the prior, in the Bayesian speak

(c) order two or pairwise

(d) $p(\mathbf{y})$ is the evidence, it appears when applying the Bayes theorem to reverse probabilities (from $p(\mathbf{x}|\mathbf{y})$ to $p(\mathbf{y}|\mathbf{x})$) and we can discard it because we are maximizing over $\mathbf{x}$ only

Problem 8

Joan Serrat *0.5 Points*

Again with respect to the binary denoising example, we saw that the final equation was

$$\arg\min_{\mathbf{x}} \sum_i \alpha x_i + \sum_{j \in \mathrm{Ne}_i} \beta x_i x_j + \sum_i \gamma x_i y_i$$

where $x_i, y_i \in \{-1, +1\}$ and Ne_i means the neighbors of pixel i. What's false then ? (can be none, one or more choices)

(a) α is related to the mean intensity we expect in a solution

(b) large and positive β makes the solution more smooth

(c) with this formulation we can express our preference for a less smooth solution around image edges

(d) γ is a parameter of the prior

(e) α, β and γ can be learned from training samples

b and c

Problem 9

Joan Serrat *0.5 Points*

Which was the solution to the three main difficulties found when trying to optimize the learn the parameters of a graphical model ? Answer writing the pairs of problem-solution labels, like "1-a, 2-b, 3-c".
Difficulties:

(1) $Z(x^i, w)$ or $\mathbb{E}_{y \sim p(y|x^i,w)} \psi(x^i, y)$ impossible to calculate

(2) N small compared to number of parameters, causing overfitting

(3) N large and therefore we have to run belief propagation N times in practice

Solutions:

(a) regularization, assuming w follows a Gaussian distribution

(b) perform stochastic gradient descent

(c) since $\psi(x, y)$ decomposes in factors, we can apply some inference method like belief propagation to compute it

1-c, 2-a, 3-b

Problem 10 Joan Serrat *0.5 Points*

Consider the graphical model of Figure 3 below where observations are binary images $16 \times 8 = 128$ pixels, that is, $x_i \in \{0, 1\}^{128}$, $y_i \in Y = \{\mathrm{a}, \mathrm{b} \dots \mathrm{z}\}$ (26 lowercase letters), $x = (x_1 \dots x_9), y = (y_1 \dots y_9)$.

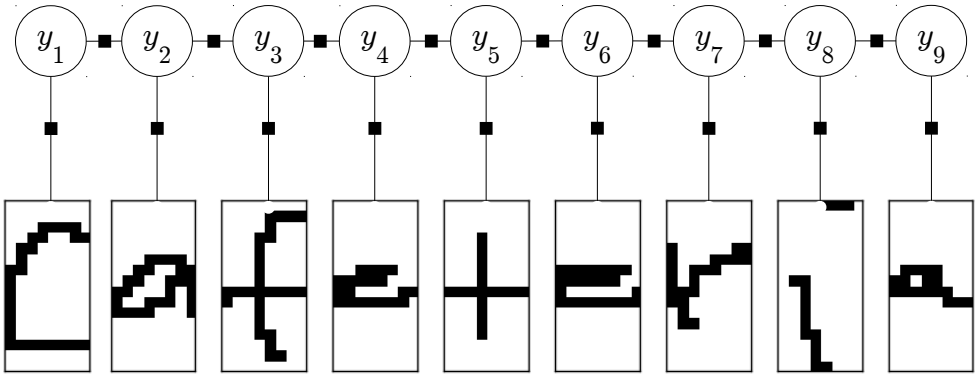


Figure 3.

We want to learn w to later infer a word from a series of binary images of letters as

$$
\begin{aligned}
y^\star &= \arg\max_{y \in \mathcal{Y}} \langle\, w, \psi(x, y) \,\rangle \\
&= \arg\max_{y \in Y^9} \sum_{i=1}^{9} \sum_{p=\mathrm{a}}^{\mathrm{z}} \sum_{j=1}^{16} \sum_{k=1}^{8} w_{pjk}\, x_{ijk} \;+\; \sum_{i=1}^{8} \sum_{p=\mathrm{a}}^{\mathrm{z}} \sum_{q=\mathrm{a}}^{\mathrm{z}} w_{pq}\, \mathbf{1}_{y_i=\mathrm{p},\, y_{i+1}=\mathrm{q}}
\end{aligned}
$$

where $\mathbf{1}_{y_i=\mathrm{p},\, y_{i+1}=\mathrm{q}}$ evaluates to 1 if $y_i = \mathrm{p}$ and $y_{i+1} = \mathrm{q}$. In this context,

(a) does it make sense to apply the technique of two stage learning ? why and how ?

(b) what does $w_{p=\mathrm{a},q=\mathrm{b}}$ mean or represent ?

(a) Yes, it makes sense to apply two-stage learning. There are about 3000 unary parameters and only 26*26 pairwise, and it is possible that the learning could be driven by the first group. Instead of one weight per pixel and possible label, we could train some classifier (like an SVM) on the binary images and use the vector of 26 probabilities as features. Then the number of unary parameters would be the same as pairwise.
(b) compatibility of 'b' right after 'a', high if this is a frequent pair in our training set

Problem 11 *Oriol Ramos Terrades, 2.5 Points*

The results obtained by the main parties the last general election draw a complex panorama where they will have to negotiate in order to chose the Prime minister of the country. According to public speeches of leaders of parties and their program, experts from the whole country have obtained the following probabilities and conditional probabilities.

First, the A party can agree with both the D party and the I party with the following *a priori* probabilities:

x_1	$p(x_1)$
Agreement with D	0.6
Agreement with I	0.2
None of them	0.2

The D Party doesn't want to reach an agreement with the A party if A agrees with the I party. The conditional probability modeling the vote of the D party is:

x_1	x_2	$p(x_2\|x_1)$
Agreement with D	Vote for A	1
Agreement with D	Vote for B	0
Agreement with D	None	0
Agreement with I	Vote for A	0.2
Agreement with I	Vote for B	0.8
Agreement with I	None	0
None	Vote for A	0.2
None	Vote for B	0.6
None	None	0.4

On the contrary, the C party wishes that the A party will agree with the I party to vote for them. The conditional probability is:

x_1	x_3	$p(x_3\|x_1)$
Agreement with D	Vote for A	0
Agreement with D	None	1
Agreement with I	Vote for A	1
Agreement with I	None	0
None	Vote for A	0.5
None	None	0.5

Moreover, the A party has a lot of problems inside their own organization and depending on the agreement they might reach with they can even vote for the leader of the B party. The conditional probability is:

x_1	x_4	$p(x_4\|x_1)$
Agreement with D	Vote for A	1
Agreement with D	Vote for B	0
Agreement with D	None	0
Agreement with I	Vote for A	0.2
Agreement with I	Vote for B	0.4
Agreement with I	None	0.4
None	Vote for A	0.2
None	Vote for B	0
None	None	0.8

Finally, the probability that they will repeat the elections will basically depend on the votes of the A and the B parties as follows:

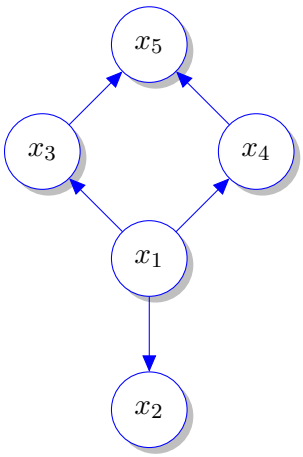
x_3	x_4	x_5	$p(x_5\|x_3, x_4)$
Vote for A	Vote for A	0	0.8
Vote for A	Vote for A	1	0.2
Vote for A	Vote for B	0	1
Vote for A	Vote for B	1	0
Vote for A	None	0	0
Vote for A	None	1	1
None	Vote for A	0	0
None	Vote for A	1	1
None	Vote for B	0	1
None	Vote for B	1	0
None	None	0	0
None	None	1	1

where $x_5 = 0$ means that parties want avoid to repeat the elections.

Assume that we have observed $x_5 = 0$ and $x_2 =$"Vote for B". We wonder if we can infer the exact probability $p(x_4|x_2, x_5)$.

a) A Bayesian net can model this problem. Draw it and write the domain of each random variable (0.5 points).

Solution:



The domain of each random variable is the following:

$x_1 = \{$'Agreement with D', 'Agreement with I', 'None of them' $\}$
$x_2 = \{$'Vote for A', 'Vote for B', 'None of them' $\}$
$x_3 = \{$'Vote for A', 'None of them' $\}$
$x_4 = \{$'Vote for A', 'Vote for B', 'None of them' $\}$
$x_5 = \{0, 1\}$

b) Write the joint distribution given by the factorization defined by the Bayesian net (0.5 points).

Solution: The joint distribution is:

$$p(x_1, x_2, x_3, x_4, x_5) = \frac{1}{Z} p(x_5|x_3, x_4)p(x_4|x_1)p(x_3|x_1)p(x_2|x_1)p(x_1) \tag{2}$$

where Z is the partition function defined as:

$$Z = \sum_{x_1,x_2,x_3,x_4,x_5} p(x_5|x_3, x_4)p(x_4|x_1)p(x_3|x_1)p(x_2|x_1)p(x_1) \tag{3}$$

c) We define the factor function, $\phi_a(x_3, x_4, x_5) = p(x_5|x_3, x_4)$, to model the interaction between these three variables. This factor defines a 3-order clique. Write the matrices $\phi_a(x_3, \cdot, x_5)$ for each possible value of x_4 (0.5 points).

solution: According to our definition of x_4, it can take 3 different values: {'Vote for A', 'Vote for B', 'None of them' }. Thus, we have 3 matrices:

$$\phi_a(x_3, \text{'Vote for A'}, x_5) = \begin{pmatrix} 0.8 & 0 \\ 0.2 & 1 \end{pmatrix} \tag{4}$$

$$\phi_a(x_3, \text{'Vote for B'}, x_5) = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \tag{5}$$

$$\phi_a(x_3, \text{'None of them'}, x_5) = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \tag{6}$$

d) Define the remaining factor functions and draw the corresponding factor graph (0.5 points)?

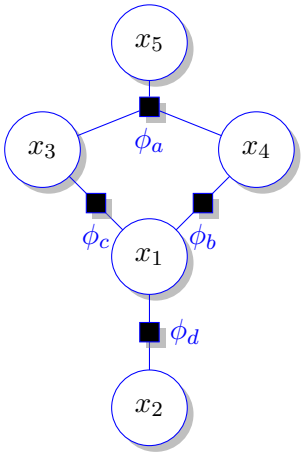
solution: We have defined the following factor functions:

$$\phi_b(x_1, x_4) = p(x_4|x_1) = \begin{pmatrix} 1 & 0.2 & 0.2 \\ 0 & 0.8 & 0.6 \\ 0 & 0 & 0.4 \end{pmatrix}$$

$$\phi_c(x_1, x_3) = p(x_3|x_1)p(x_1) = \begin{pmatrix} 0 & 1 & 0.5 \\ 1 & 0 & 0.5 \end{pmatrix} \circ \begin{pmatrix} 0.6 & 0.2 & 0.2 \\ 0.6 & 0.2 & 0.2 \end{pmatrix} = \begin{pmatrix} 0 & 0.2 & 0.1 \\ 0.6 & 0 & 0.1 \end{pmatrix}$$

$$\phi_d(x_1, x_2) = p(x_2|x_1) = \begin{pmatrix} 1 & 0.2 & 0.2 \\ 0 & 0.4 & 0 \\ 0 & 0.4 & 0.8 \end{pmatrix}$$

With these factor functions, the factor graph is:



e) Does Belief Propagation (BP) provide exact estimates of marginals? compute the message $m_{3\leftarrow d}(x_3)$ if it provides exact inference. Otherwise explain why you can not apply BP (0.5 points).

solution: Although the Bayesian network is an acyclic graph and it factorizes, the corresponding factor graph has a loop that make unfeasible to apply BP for exact inference. Loopy Belief Propagation can applied instead but we cannot guarantee its convergence.

In particular, to send a message from factor c to x_3, x_1 has to receive before messages from factors d and b. x_2 has been observed, so d can send its message to x_1 without any problem. However, x_4 has to receive a message from x_5 and x_3, which is impossible because x_3 is hidden.