Machine Learning: Exercises for Block III

Isabel Valera

Exercise 1: Lagrange Dual Problem

Consider the optimization problem with variable $x \in \mathbb{R}$ given in equation 1.

i) State the dual problem and verify that it is a concave maximization problem.

ii) Find the dual optimal value and dual optimal solution λ^*.

iii) Does strong duality hold?

$$\begin{aligned} &\text{minimize } x^2 + 1 \\ &\text{subject to } (x-2)(x-4) \leq 0 \end{aligned} \tag{1}$$

Exercise 2: Inequality constraint

i) Express the dual problem of the primal problem given in 2 with $c \neq 0$ in terms of the conjugate f^*.

ii) Explain why the dual problem you give is convex. We do not assume f is convex.

$$\begin{aligned} &\text{minimize } c^\top x \\ &\text{subject to } f(x) \leq 0 \end{aligned} \tag{2}$$

Exercise 3: KKT conditions

i) Derive the KKT conditions for the problem given in problem 3 with variable $\boldsymbol{X} \in \boldsymbol{S}^n$ (n dimensional symmetric) and domain $\boldsymbol{S}^n_{++}$ (symmetric positive-definite). $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{s} \in \mathbb{R}^n$ are given with $\boldsymbol{s}^\top \boldsymbol{y} = 1$.

ii) Verify that the optimal solution is given by equation 4.

$$\begin{aligned} &\text{minimize } \mathbf{tr}(\boldsymbol{X}) - \log\det \boldsymbol{X} \\ &\text{subject to } \boldsymbol{X}\boldsymbol{s} = \boldsymbol{y} \end{aligned} \tag{3}$$

$$\boldsymbol{X}^* = \mathbb{I} + \boldsymbol{y}\boldsymbol{y}^\top - \frac{1}{\boldsymbol{s}^\top \boldsymbol{s}} \boldsymbol{s}\boldsymbol{s}^\top \tag{4}$$

Exercise 4: Estimating covariance and mean

We consider the problem of estimating the covariance matrix $\boldsymbol{\Sigma}$ and the mean $\boldsymbol{\mu}$ of a Gaussian probability density function as given in equation 5 based on N independent samples $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N \in \mathbb{R}^n$.

i) We first consider the estimation problem when there are no additional constrains on $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$. Let $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ be the sample mean and covariance as defined in equations 6. Show that the log-likelihood function given in equation 7 can be expressed as in equation 8 and use this expression to show that if $\hat{\boldsymbol{\Sigma}} \succ 0$ the ML estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ are unique and given by the sample covariance and sample mean.

ii) The log-likelihood function includes a convex term $(-\log\det \boldsymbol{\Sigma})$ so it is not obviously concave. Show that $\mathcal{L}$ is concave, jointly in $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ in the region defined by $\boldsymbol{\Sigma} \preceq 2\hat{\boldsymbol{\Sigma}}$. This means we can use convex optimization to compute simultaneous ML estimates of $\boldsymbol{\Sigma}$ and μ, subject to convex constraints, as long as the constraints include $\boldsymbol{\Sigma} \preceq 2\hat{\boldsymbol{\Sigma}}$, i.e. the estimate $\boldsymbol{\Sigma}$ must not exceed twice the unconstrained ML estimate.

$$p(\boldsymbol{x} \mid \boldsymbol{\Sigma}, \boldsymbol{\mu}) = (2\pi)^{-\frac{n}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{5}$$

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_k \\ \hat{\boldsymbol{\Sigma}} &= \frac{1}{N}\sum_{k=1}^{N} (\boldsymbol{x}-\hat{\boldsymbol{\mu}})(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^\top \end{aligned} \tag{6}$$

$$\mathcal{L}(\boldsymbol{\Sigma}, \boldsymbol{\mu}) = -\frac{Nn}{2}\log(2\pi) - \frac{N}{2}\log\det \boldsymbol{\Sigma} - \frac{1}{2}\sum_{k=1}^{N}(\boldsymbol{x}_k-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_k-\boldsymbol{\mu}) \tag{7}$$

$$\mathcal{L}(\boldsymbol{\Sigma}, \boldsymbol{\mu}) = \frac{N}{2}\left(-n\log(2\pi) - \log\det \boldsymbol{\Sigma} - \mathbf{tr}(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}) - (\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\right) \tag{8}$$

Exercise 5: Estimating mean and variance

Consider a random variable $x \in \mathbb{R}$ with density p, which is normalized, i.e. has zero mean and unit variance. Consider a random variable $y = \frac{x+b}{a}$ obtained by an affine transformation of x, where $a > 0$. The random variable y has mean $\frac{b}{a}$ and variance $\frac{1}{a^2}$. As a and b vary over the non-negative real numbers $\mathbb{R}_+$ and the real numbers $\mathbb{R}$, respectively, we generate a family of densities obtained from p by scaling and shifting, uniquely parametrized by mean and variance.

i) Show that if p is log-concave, then finding the ML estimate of a and b, given samples $y_1, ..., y_n$ of y is a convex problem.

ii) As an example, work out an analytical solution for the ML estimates of a and b, assuming p is a normalized Laplacian density $p(x) = \exp(-2|x|)$.

Exercise 6: Robust linear classification

Consider the robust linear classification problem given in problem 9 where we seek an affine function $f(x) = \boldsymbol{w}^\top \boldsymbol{x} - b$ that separates the two sets of points $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ and $\{\boldsymbol{y}_1, ..., \boldsymbol{y}_M\}$. This means that $\boldsymbol{w}^\top \boldsymbol{x}_i - b > 0$ for $i = 1, ..., N$ and $\boldsymbol{w}^\top \boldsymbol{y}_j - b < 0$ for $j = 1, ..., M$.

i) Show that the optimal value t^* is positive if and only if the two sets of points can be linearly separated. When the two sets of points can be linearly separated, show that the inequality $||\boldsymbol{w}||_2 \leq 1$ is tight, i.e., we have $||\boldsymbol{w}^*||_2 = 1$ for the optimal $\boldsymbol{w}^*$.

ii) Using the change of variables $\tilde{\boldsymbol{w}} = \frac{\boldsymbol{w}}{t}, \tilde{b} = \frac{b}{t}$, prove that problem 9 is equivalent to the quadratic program given in 10.

$$\begin{aligned} \text{maximize } & t \\ \text{subject to } & \boldsymbol{w}^\top \boldsymbol{x}_i - b \geq t, \ i = 1, ..., N \\ & \boldsymbol{w}^\top \boldsymbol{y}_i - b \leq -t, \ i = 1, ..., M \\ & ||\boldsymbol{w}||_2 \leq 1 \end{aligned} \tag{9}$$

$$\begin{aligned} \text{minimize } & ||\tilde{\boldsymbol{w}}||_2 \\ \text{subject to } & \tilde{\boldsymbol{w}}^\top \boldsymbol{x}_i - \tilde{b} \geq 1, \ i = 1, ..., N \\ & \tilde{\boldsymbol{w}}^\top \boldsymbol{y}_i - \tilde{b} \leq -1, \ i = 1, ..., M \end{aligned} \tag{10}$$

Exercise 7: Linear discrimination and weight errors

Suppose we are given two sets of points $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ and $\{\boldsymbol{y}_1, ..., \boldsymbol{y}_M\}$ in $\mathbb{R}^n$ that can be linearly separated. As in problem 9 we can find the affine function that discriminates the sets, and gives the largest gap in function values. We can also consider robustness with respect to changes in the weight vector $\boldsymbol{w}$. For a given $\boldsymbol{w}$ and b for which $f(x) = \boldsymbol{w}^\top \boldsymbol{x} - b$ separates the two sets, we define the weight error margin as the norm of the smallest $\boldsymbol{u} \in \mathbb{R}^n$ such that the affine function $(\boldsymbol{w} + \boldsymbol{u})^\top \boldsymbol{x} - b$ no longer separates the two sets of points. In other words, the weight error margin is the maximum ρ such that $(\boldsymbol{w} + \boldsymbol{u})^\top \boldsymbol{x}_i \geq b$ for $i = 1, ..., N$ and $(\boldsymbol{w} + \boldsymbol{u})^\top \boldsymbol{y}_j \leq b$ for $j = 1, ..., M$ holds for all $\boldsymbol{u}$ with $||\boldsymbol{u}||_2 \leq \rho$. Show how to find $\boldsymbol{w}$ and b that maximize the weight error margin, subject to the normalization constraint $||\boldsymbol{w}||_2 \leq 1$.

References

[1] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.