MIDTERM EXAM RESULTS

Average : 89 (94)

Stdev : 7

$$w^{k+1} = w^k - \alpha \nabla f(w^k)$$

$$\alpha (A w^k - b)$$

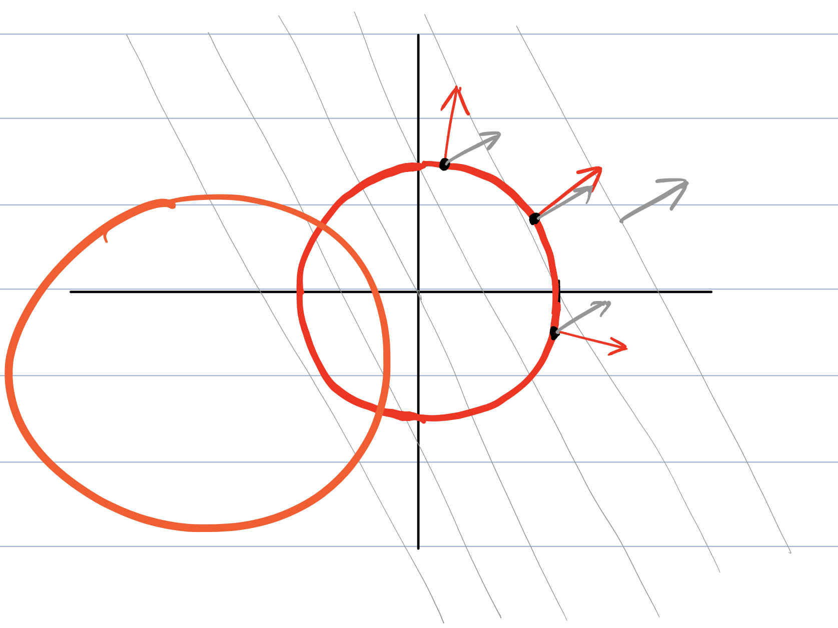
LAGRANGE MULTIPLIERS

Constrained optimization is harder than unconstrained optimization, so if we can convert the latter into the former, that's great.

BASIC EXAMPLE

Consider the following basic problem:

$$\underset{x,y}{\operatorname{argmax}} \quad 2x + y$$
$$\text{subject to} \quad x^2 + y^2 = 1$$



Consider the geometric intuition: can you move along the constraint surface to improve the objective? When do you reach an extremum?

Method of Lagrange multipliers:

- Convert constraint surface to implicit surface $g(x)=0$
- At the extrema points, gradients are collinear, which means that

$$\nabla_x f = -\lambda \nabla_x g$$

$$\nabla(f+\lambda g) = 0$$

So if we have a function $L(x,\lambda) = f(x) + \lambda g(x)$, then the unconstrained extrema of L correspond to the extrema of f constrained to g.

In our example setting,

$$f(x,y) = 2x + y$$
$$g(x,y) = x^2 + y^2 - 1$$

Our "Lagrangian" is:

$$L(x,y,\lambda) = 2x + y + \lambda(x^2 + y^2 - 1)$$

$$\frac{\partial L}{\partial x} = 2 + 2\lambda x = 0 \qquad \lambda x = -1 \qquad x = \frac{-1}{\lambda}$$

$$\frac{\partial L}{\partial y} = 1 + 2\lambda y = 0 \qquad \lambda y = -1/2 \qquad y = \frac{-1}{2} \cdot \frac{1}{\lambda}$$

$$x^2 + y^2 = 1$$

$$\frac{-1}{\lambda^2} - \frac{1}{2\lambda^2} = 1$$

$$-\frac{2}{2\lambda^2} - \frac{1}{2\lambda^2} = 1$$

$$-3 = 2\lambda^2$$

$$\lambda = \left(-\frac{3}{2}\right)^{1/2} \qquad x = \frac{1}{\sqrt{3/2}} \qquad y = \frac{1}{2\sqrt{3/2}}$$

Now we can use this same idea* on SVMs:

$$\min_{w,b} \; \frac{1}{2}\|w\|^2 + C \sum_i \xi_i$$

$$\text{s.t.} \quad y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

We add one new variable for each constraint, which we call α_i and β_i:

$$L(w, b, \vec{\xi}, \vec{\alpha}, \vec{\beta}) =$$

$$\frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \beta_i \xi_i - \sum_i \alpha_i \Big(\big(y_i \langle w, x_i \rangle + b \big) - 1 + \xi_i \Big)$$

We need to be careful about the objective: it's

$$\min_{w,b,\xi} \; \max_{\alpha \geq 0} \; \max_{\beta \geq 0} \; L(w, b, \vec{\xi}, \vec{\alpha}, \vec{\beta})$$

Now we start massaging that expression. We are interested in Kernelization, so we are looking to remove w from the equation.

Remember that the method of L.M. asks us to take derivatives wrt the variables of interest and set them to zero. Let's do that with w:

$$\nabla_w L = w - \sum_i \alpha_i y_i x_i = 0$$

$$w = \sum_i \alpha_i y_i x_i \quad \leftarrow \text{representer!}$$

Now substitute w back in L:

$$L(b, \xi, \alpha, \beta) =$$

$$\frac{1}{2} \left\| \sum_j \alpha_j y_j x_j \right\|^2 + C \sum_i \xi_i - \sum_i \beta_i \xi_i$$
$$- \sum_i \alpha_i \left(y_i \left(\left\langle \sum_j \alpha_j y_j x_j, x_i \right\rangle + b \right) - 1 + \xi_i \right)$$

Now we use $\|x\|^2 = \langle x, x \rangle$ and $\left\langle \sum_i v_i, x \right\rangle = \sum \langle v_i, x \rangle$:

$$L(b, \xi, \alpha, \beta) = \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$
$$+ \sum_i (C - \beta_i) \xi_i$$
$$- \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$
$$- \sum_i \alpha_i \left(y_i b - 1 + \xi_i \right)$$

$$= -\frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \qquad (1)$$

$$+ \sum_i (C - \beta_i)\, \xi_i \qquad (2)$$

$$- b \sum_i \alpha_i y_i \qquad (3)$$

$$- \sum_i \alpha_i (\xi_i - 1) \qquad (4)$$

Now each of these terms simplify with knowledge of ∇L.

$$\frac{\partial L}{\partial b} = -\sum \alpha_i y_i = 0$$

this means (3) is zero at extremum!

$$\frac{\partial L}{\partial \xi_i} = (C - \beta_i) - \alpha_i = 0$$

this means β_i disappear (except for new constraint)

$$C - \beta_i = \alpha_i$$

$$= -\frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \qquad (1)$$

$$+ \sum \cancel{(C - \beta_i)\, \xi_i} \quad \cancel{\sum_i \alpha_i \xi_i} \qquad (2)$$

$$\cancel{- b \sum_i \alpha_i y_i} \qquad (3)$$

these cancel!

$$- \sum_i \alpha_i (\cancel{\xi_i} - 1) \qquad (4)$$

$$L(\alpha) = -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle + \sum_i \alpha_i \quad (!!)$$

subject to $0 \leq \alpha_i \leq C$ $\qquad w = \sum \underline{\underline{\alpha_i}} y_i x_i$

In matrix form this is even cleaner:

$$L(\vec{\alpha}) = \langle \vec{\alpha}, \vec{1} \rangle - \frac{1}{2} \vec{\alpha}^T G \alpha$$

$$G = \left(y_i \cdot y_j \langle x_i, x_j \rangle \right)_{ij} \leftarrow \text{Kernel!}$$

$$\nabla_\alpha L = \vec{1} - G\vec{\alpha}$$

To optimize this, use projected gradient descent: after each step, check if variables went outside feasible region, then project them back.

PREDICTION ON KERNELIZED SVM

Without Kernels,

$$f(\hat{x}) = \text{sign}(\langle w, \hat{x}\rangle + b)$$

With Kernels:

representer!

$$f(\hat{x}) = \text{sign}\left(\left\langle \sum \alpha_i y_i x_i, \hat{x}\right\rangle\right)$$

$$= \text{sign}\left(\sum_i \alpha_i y_i \langle \hat{x}, x_i\rangle\right)$$

If $\alpha_i = 0$, then x_i has no influence in prediction!

Notice that the α_i were the variables we added in the Lagrange formulation. These are known as the dual variables, and the optimization problem written entirely on dual variables is called the dual problem.