UMC 203: AI and ML

Yash Kamble

March 2025

Contents

Lectures

Chapter I

Convex Optimisation

Lecture 1. Sunday March 02

Definition I.1 (Convex function)**.** A set $C \subseteq \mathbb{R}^d$ is said to be *convex* if for all $x, y \in C$ and $\lambda \in [0, 1]$,

$$(1 - \lambda)x + \lambda y \in C.$$

A function $f : C \to \mathbb{R}$ over a convex set $C \subseteq \mathbb{R}^d$ is said to be *convex* if for all $x, y \in C$ and $\lambda \in [0, 1]$,

$$f((1 - \lambda)x + \lambda y) \le (1 - \lambda)f(x) + \lambda f(y).$$

Fact I.2. *Let $f \in C^1(C)$, where $C \subseteq \mathbb{R}^d$ is convex. Then f is convex iff*

$$f(y) - f(x) \ge \langle \nabla f(x),\, y - x \rangle$$

for all $x, y \in C$.

Notation. Let A and B be symmetric matrices. We write $A \succeq B$ if $A - B$ is positive semidefinite.

Proposition I.3. *$\succeq$ is a partial order.*

Proof.

- Reflexivity: $A - A = 0 \succeq 0$.
- Antisymmetry: $A - B \succeq 0$ and $B - A \succeq 0$ implies $A - B = 0$, since if λ is an eigenvalue of $A - B$, then $-\lambda$ is an eigenvalue of $B - A$. But all eigenvalues of $A - B$ as well as $B - A$ are nonnegative, so $\lambda = 0$.

- Transitivity: Suppose $A \succeq B \succeq C$. Then for all u,

$$\begin{aligned} &\langle u,\, (A-B)u\rangle \geq 0 \\ &\langle u,\, (B-C)u\rangle \geq 0 \\ \implies &\langle u,\, (A-C)u\rangle = \langle u,\, (A-B+B-C)u\rangle \\ &\qquad = \langle u,\, (A-B)u\rangle + \langle u,\, (B-C)u\rangle \\ &\qquad \geq 0. \end{aligned}$$

□

Fact I.4. *Let $f \in C^2(C)$, where $C \subseteq \mathbb{R}^d$ is convex. Let $H(x) = (\operatorname{Hess} f)(x)$. Then f is convex iff*

$$H(x) \succeq 0 \quad \forall x \in C.$$

Definition I.5 (Convex optimisation problem)**.** Let $f\colon \mathbb{R}^d \to \mathbb{R}$ and $f_i\colon \mathbb{R}^d \to \mathbb{R}$ be convex functions for each $i \in [m]$. Let $(a_j)_{j=1}^n \subseteq \mathbb{R}^d$ and $(b_j)_{j=1}^n \subseteq \mathbb{R}$. The *convex optimisation problem* is to find

$$\min_{x\in\mathbb{R}^d} f(x) \quad \text{such that} \quad \begin{cases} f_i(x) \leq 0 \text{ for all } i \in [m], \\ \langle a_j,\, x\rangle = b_j \text{ for all } j \in [n]. \end{cases}$$

I.1 KKT Conditions

Definition I.6. Let $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n$. The *Lagrangian* of the convex optimisation problem is

$$L(x,\lambda,\mu) = f(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{j=1}^n \mu_j(\langle a_j,\, x\rangle - b_j).$$

We say that x^* is a *KKT point* if there exist λ and μ such that

$$\begin{aligned} \nabla_x L(x^*,\lambda,\mu) &= 0, \\ \langle a_j,\, x^*\rangle - b_j &= 0 \quad \forall j \in [n], \\ f_i(x^*) &\leq 0 \quad \forall i \in [m], \\ \lambda_i f_i(x^*) &= 0 \quad \forall i \in [m]. \end{aligned}$$

The first condition is the *stationarity* condition. The second and third conditions are the *primal feasibility* conditions. The final condition is the *complementary slackness* condition.

Fact I.7. *If x^* is a KKT point for the convex optimisation problem, then x^* is a global minimiser.*

Example. Consider the convex optimisation problem

$$\min_{x \in \mathbb{R}^d} \frac{1}{2}\|x - z\|^2 \quad \text{such that} \quad \langle w, x \rangle + b = 0.$$

The Lagrangian is

$$L(x, \mu) = \frac{1}{2}\|x - z\|^2 + \mu(\langle w, x \rangle + b).$$

The KKT conditions are

$$\begin{aligned} \nabla_x L(x^*, \mu) = x - z + \mu w = 0, \\ \implies x^* = z - \mu w, \\ \langle w, x^* \rangle + b = 0 \\ \implies \langle w, z - \mu w \rangle + b = 0 \\ \implies \langle w, z \rangle - \mu\|w\|^2 + b = 0 \end{aligned}$$

So the minimizer is

$$x^* = z - \frac{(\langle w, z \rangle + b)}{\|w\|^2} w.$$

This is the orthogonal projection of z onto the hyperplane.

I.2 Wolfe Dual

Definition I.8 (Wolfe dual)**.** For a given convex optimisation problem P, the *Wolfe dual* problem D is

$$\max_{x,\lambda,\mu} L(x, \lambda, \mu) \quad \text{such that} \quad \begin{cases} \lambda \geq 0, \\ \nabla_x L(x, \lambda, \mu) = 0. \end{cases}$$

Theorem I.9. *If x^* is a KKT point for the convex optimisation problem with Lagrange multipliers λ^* and μ^*, then (x^*, λ^*, μ^*) solves the Wolfe dual problem.*

Proof. First absorb the affine equality constraints into the convex inequality constraints. Suppose x^* is a KKT point with Lagrange multipliers λ^*. Note that (x^*, λ^*) is feasible for the Wolfe dual problem. Then

$$\begin{aligned} L(x^*, \lambda^*) &= f(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) \\ &= f(x^*) \end{aligned}$$

by complementary slackness. Also note that by primal feasibility,

$$\begin{aligned} L(x^*, \lambda) &= f(x^*) + \sum_{i=1}^m \lambda_i f_i(x^*) \\ &\le f(x^*) = L(x^*, \lambda^*). \end{aligned}$$

Let $f_0 = f$. Now since f_i, $i \in \{0, \dots, m\}$, are convex, we have

$$f_i(x^*) \ge f_i(x) + \langle \nabla f_i(x),\, x^* - x \rangle$$

for all x.

Thus

$$\begin{aligned} L(x^*, \lambda) &= f(x^*) + \sum_{i=1}^m \lambda_i f_i(x^*) \\ &\ge f(x) + \langle \nabla f(x),\, x^* - x \rangle + \sum_{i=1}^m \lambda_i (f_i(x) + \langle \nabla f_i(x),\, x^* - x \rangle) \\ &= f(x) + \sum_{i=1}^m \lambda_i f_i(x) + \left\langle \nabla f(x) + \sum_{i=1}^m \lambda_i \nabla f_i(x),\, x^* - x \right\rangle \\ &= L(x, \lambda) + \langle x^* - x,\, \nabla_x L(x, \lambda) \rangle. \end{aligned}$$

Then if x is a feasible point for the Wolfe dual problem,

$$L(x^*, \lambda) \ge L(x, \lambda)$$

by the stationarity condition. Thus for all feasible x and λ,

$$L(x^*, \lambda^*) \ge L(x^*, \lambda) \ge L(x, \lambda). \qquad \square$$

Thus we can use the Wolfe dual to hunt for KKT points.

Chapter II

Large margin classification

Lecture 2.
Sunday
March 02

Let $\mathcal{D} = \left\{(x^{(i)}, y_i)\right\}_{i=1}^{m}$ be a linearly separable dataset. The perceptron algorithm finds a separating hyperplane, but there are many such hyperplanes. Which one is the best?

We can focus on the margin of the hyperplane. The margin is as defined in **??**. The hyperplane with the largest margin is deemed the best.

Definition II.1 (The SVM problem)**.** The *support vector machine* (SVM) problem is to find the hyperplane with the largest margin. That is, find w that solves

$$\max_w \min_i \frac{y_i\langle w,\, x^{(i)}\rangle}{\|w\|}.$$

What about the more general classifiers using $\langle w,\, x\rangle + b$? We can append a constant 1 to each $x^{(i)}$ and append b to w. Hence we can restrict our attention to the case where $b = 0$.

Note that the objective function is homogeneous in w. So we can scale w such that $\min_i y_i\langle w,\, x^{(i)}\rangle = 1$. Then the problem becomes

$$\max_w \frac{1}{\|w\|} \quad \text{subject to} \quad \min_i y_i\langle w,\, x^{(i)}\rangle = 1.$$

When is $\min_i y_i\langle w,\, x^{(i)}\rangle = 1$? When $\langle w,\, y_i x^{(i)}\rangle \geq 1$ for all i, but also $\langle w,\, y_i x^{(i)}\rangle = 1$ for some i. What if $\langle w,\, y_i x^{(i)}\rangle > 1$ for all i? Then w can be shrunken to increase the objective while still satisfying the constraints. Thus the problem becomes

$$\max_w \frac{1}{\|w\|} \quad \text{subject to} \quad \langle w,\, y_i x^{(i)}\rangle \geq 1 \text{ for all } i.$$

But maximizing $1/\|w\|$ is the same as minimizing $\|w\|^2$. So we again rewrite the problem as

$$\min_w \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad \langle w,\, y_i x^{(i)}\rangle \geq 1 \text{ for all } i.$$

Note that $w \mapsto \|w\|^2$ is a strictly convex function. We have the Lagrangian

$$L(w,\lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^m \lambda_i(\langle w,\, y_i x^{(i)}\rangle - 1)$$

and so the KKT conditions

$$\nabla_w L(w,\lambda) = 0 \implies w = \sum_{i=1}^m \lambda_i y_i x^{(i)} \tag{II.1}$$

$$\langle w,\, y_i x^{(i)}\rangle \geq 1 \quad \text{for all } i, \tag{II.2}$$

$$\lambda_i(\langle w,\, y_i x^{(i)}\rangle - 1) = 0 \quad \text{for all } i. \tag{II.3}$$

If $\lambda_i > 0$, then $\langle w,\, y_i x^{(i)}\rangle = 1$. If $\langle w,\, y_i x^{(i)}\rangle > 1$, then $\lambda_i = 0$.

The $x^{(i)}$s for which $\lambda_i > 0$ are called the *support vectors*. These are at most the points for which $\langle w,\, y_i x^{(i)}\rangle = 1$.

Lecture 3. Sunday March 02

Substituting equation (II.1) into the Lagrangian gives

$$\begin{aligned} L &= \frac{1}{2}\Big\|\sum_i \lambda_i y_i x^{(i)}\Big\|^2 - \sum_i \lambda_i \Big\langle \sum_j \lambda_j y_j x^{(j)},\, y_i x^{(i)}\Big\rangle + \sum_{i=1}^m \lambda_i \\ &= \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j \langle y_i x^{(i)},\, y_j x^{(j)}\rangle \end{aligned}$$

Thus using the Wolfe dual (section I.2), the SVM problem is to solve

$$\max_\lambda \sum_i \lambda_i - \frac{1}{2}\sum_{i,j}\lambda_i\lambda_j \Big\langle y_i x^{(i)},\, y_j x^{(j)}\Big\rangle \quad \text{subject to} \quad \lambda_i \geq 0$$

If we find such a λ, we have

$$w = \sum_{i=1}^m \lambda_i y_i x^{(i)}$$

and the classifier

$$h(x) = \operatorname{sgn}\langle w,\, x\rangle.$$

Except… this is **NOT** the SVM problem. The SVM problem does not absorb the constant b into the vector w.

Definition II.1 (The SVM problem)**.** The *support vector machine* (SVM) problem is to find the hyperplane with the largest margin. That is, find w and b that solve

$$\max_{w,b} \min_i \frac{y_i(\langle w,\, x^{(i)}\rangle + b)}{\|w\|}.$$

Notice that the norm in the denominator *does not* include b. Through much the same machinery as before, one arrives at the problem

$$\boxed{\max_\lambda \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i\lambda_j y_i y_j \langle x^{(i)},\, x^{(j)}\rangle \quad \text{subject to} \quad \begin{cases}\lambda_i \geq 0, \\ \sum_i \lambda_i y_i = 0.\end{cases}}$$

This determines w as before: $w = \sum_i \lambda_i y_i x^{(i)}$.

Note that the only dependence on $x^{(i)}$ is through the inner product. Thus we can use the *kernel trick* to solve the SVM problem in linearly non-separable cases.

Suppose $(x^{(i)}, y_i)$ are not linearly separable, but there is a transformation Φ such that $(\Phi(x^{(i)}), y_i)$ are linearly separable. Then we can apply SVM to the transformed dataset.

$$\max_\lambda \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i\lambda_j \Big\langle y_i\Phi(x^{(i)}),\, y_j\Phi(x^{(j)})\Big\rangle \quad \text{subject to} \quad \begin{cases}\lambda_i \geq 0, \\ \sum_i \lambda_i y_i = 0.\end{cases}$$

If we can compute $\Big\langle \Phi(x^{(i)}),\, \Phi(x^{(j)})\Big\rangle$, then we can solve the SVM problem for the transformed dataset.

II.1 Generalization Error

Lecture 4. Sunday March 02

Theorem II.2 (FOML 5.4)**.** *Let $h_\mathcal{D}^{SVM}$ be the classifier returned by the SVM for a sample $\mathcal{D}$, and let $N_{SV}(\mathcal{D})$ be the number of support vectors that define $h_\mathcal{D}^{SVM}$. Then,*

$$\mathop{\boldsymbol{E}}_{\mathcal{D}\sim P^m}(R(h_\mathcal{D}^{SVM})) \leq \mathop{\boldsymbol{E}}_{\mathcal{D}\sim P^{m+1}}\left[\frac{N_{SV}(\mathcal{D})}{m+1}\right]$$

Proof. The proof is identical to that of **??**, proceeding via the leave-one-out error. □

If the training set error is zero, is the generalization error also zero?

Fact II.3. *Let $Z_1, \ldots, Z_n$ be iid random variables with* $\mathrm{P}(Z_i \in [a,b]) = 1$ *and* $\boldsymbol{E}[Z_i] = \mu$.

Then,

$$\mathrm{P}\Big(\big|\overline{Z}-\mu\big|\geq\epsilon\Big)\leq 2\exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)$$

where $\overline{Z}=\frac{1}{n}\sum_{i=1}^{n}Z_i$.

We can use this to give probabilistic bounds on the generalization error using its expected value (since it is bounded between 0 and 1).

II.2 VC Dimension

Let $\mathcal{H}$ be a hypothesis class. That is, a set of functions from $\mathcal{X}$ to $\mathcal{Y}$. For our purposes, $\mathcal{Y}=\{-1,1\}$.

Definition II.4 (Growth function)**.** The growth function $\Pi_{\mathcal{H}}\colon\mathbb{N}\to\mathbb{N}$ is defined by

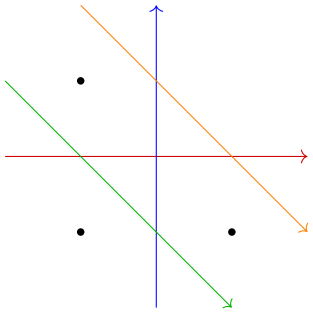
$$\Pi_{\mathcal{H}}(m)=\max_{x_1,\ldots,x_m\in\mathcal{X}}\#\{(h(x_1),\ldots,h(x_m))\mid h\in\mathcal{H}\}$$

In other words, $\Pi_{\mathcal{H}}(m)$ is the maximum number of distinct ways in which m points can be classified by functions in $\mathcal{H}$.

Notation. We will denote the set of affine classifiers from $\mathbb{R}^d$ to $\{-1,1\}$ by $\mathcal{L}_d$.

Example. If $\mathcal{H}=\mathcal{L}_2$, then

$$\begin{aligned}\Pi_{\mathcal{H}}(1)&=2\\ \Pi_{\mathcal{H}}(2)&=4\\ \Pi_{\mathcal{H}}(3)&=8\end{aligned}$$



These four classifiers give four distinct ways to classify the given three points. Reversing these gives another four ways. There are only eight possible labelings, so $\Pi_{\mathcal{H}}(3)=8$.

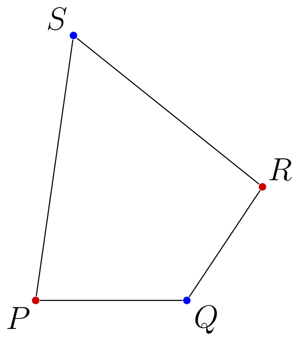
However, $\Pi_{\mathcal{H}}(4) < 16$. That is, no matter which 4 points we choose, we can't find 16 distinct classifications of them by functions in $\mathcal{H}$. In other words, for any 4 points, there exists a labeling of them that cannot be achieved by any function in $\mathcal{H}$.

Theorem II.5. $\Pi_{\mathcal{L}_2}(4) < 16$.

Proof. Let P, Q, R and S be any four points, colored red or blue. The key observation is that if a line L separates the red points from the blue points, then for any two points A and B, L intersects the line segment $\overline{AB}$ iff A and B are colored differently.

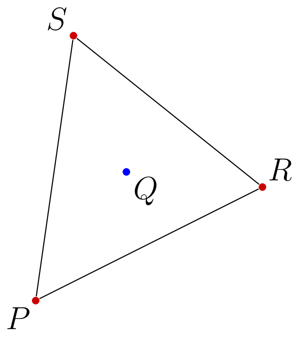
If any three points P, Q and R are collinear in that order, color them red, blue and red respectively. Then any line separating the red points from the blue points must pass through both $\overline{PQ}$ and $\overline{QR}$. The only line that does this is the line $\overline{PR}$ itself, which will assign the same color to each of these.

Now suppose that P, Q, R and S are such that no three are collinear. If they form a convex quadrilateral, color them alternately red and blue.



Any line separating the red points from the blue points must intersect every side of the quadrilateral, which is not possible.

If they form a non-convex quadrilateral, the convex hull must be a triangle. Color the points of the triangle red, and the interior point blue.



A separating plane can pass through none of the sides of the triangle, so it cannot enter the interior of the triangle at all, and thus cannot separate Q from the other points. □

Definition II.6 (Shattering)**.** A hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is said to *shatter* a set $C \subseteq \mathcal{X}$ if for every labelling of C by $\mathcal{Y}$, there exists a function $h \in \mathcal{H}$ that achieves that labelling. That is,

$$\forall y \in \mathcal{Y}^C \; \exists h \in \mathcal{H} \; \forall x \in C(h(x) = y(x))$$

Example. From the above theorem, we conclude that $\mathcal{L}_2$ shatters no set of four points.

From the example preceding it, we conclude that the set of linear classifiers in $\mathbb{R}^2$ shatters that particular set of three points (and indeed, any set of three points that are not collinear).

Definition II.7 (VC-dimension)**.** The *VC-dimension* of a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is the size of the largest set that can be shattered by $\mathcal{H}$. That is,

$$\mathrm{VC}(\mathcal{H}) = \max\{m \mid \Pi_{\mathcal{H}}(m) = |\mathcal{Y}|^m\}$$

Example. The VC-dimension of the set of linear classifiers in $\mathbb{R}^2$ is 3. This is because it shatters at least one set of three points, but no set of four points.

Theorem II.8. *The VC-dimension of the set of linear classifiers from $\mathbb{R}^d$ to $\{-1, 1\}$ is $d + 1$.*

Proof. Induction. For $d = 1$, the points -1 and 1 can obviously be shattered, using the affine maps $x \mapsto x$, $x \mapsto -x$, $x \mapsto x + 2$ and $x \mapsto -x - 2$.

Also, any three points are collinear, so they cannot be shattered by the same argument as in the proof of theorem II.5. Thus $\mathrm{VC}(\mathcal{L}_1) = 2$.

Suppose that $\mathrm{VC}(\mathcal{L}_{d-1}) = d$. Then let $P_1, \ldots, P_d$ be a set shattered by $\mathcal{L}_{d-1}$. Let $Q_i = (P_i, 0)$ for $i = 1, \ldots, d$, and $Q_{d+1} = (0, \ldots, 0, 1)$. We claim that the set $\{Q_1, \ldots, Q_{d+1}\}$ can be shattered by $\mathcal{L}_d$.

Fix a coloring y of $\{Q_1, \ldots, Q_{d+1}\}$. Consider the same coloring applied to $\{P_1, \ldots, P_d\}$ (each P_i colored the same as Q_i). Let $h(x) = \mathrm{sgn}(\langle w, x\rangle + b)$ be the classifier that achieves this coloring. WLOG assume that Q_{d+1} is colored $+1$. Let $w' = (w, 1-b)$. Then $h(x) = \mathrm{sgn}(\langle w', x\rangle + b)$ achieves the coloring y of $\{Q_1, \ldots, Q_{d+1}\}$. Thus $\mathrm{VC}(\mathcal{L}_d) \geq d + 1$.

Lecture 4: Linear SVM classifiers for linearly non-separable data: VC Dimension

To show that $\mathrm{VC}(\mathcal{L}_d) < d+2$, consider any set of $d+2$ points in $\mathbb{R}^d$. Suppose that they are shattered by $\mathcal{L}_d$. Fix a coloring y of these points. Consider the same coloring applied to any $d+1$ points, viewed as points in $\mathbb{R}^{d-1}$. Since there exists a classifier that achieves this coloring in $\mathbb{R}^d$, its restriction to $\mathbb{R}^{d-1}$ achieves the same coloring in $\mathbb{R}^{d-1}$. But this is impossible, since $\mathrm{VC}(\mathcal{L}_{d-1}) = d$. Thus $\mathrm{VC}(\mathcal{L}_d) < d+2$.

Winduction. □

Fact II.9. *Let*

$$\mathcal{X} = \left\{x \in \mathbb{R}^d \mid \|x\| \le R\right\}, \text{ and}$$
$$\mathcal{H}_B = \left\{h \in \{-1,1\}^{\mathcal{X}} \mid h = \operatorname{sgn}(\langle w,\, \cdot\rangle + b) \text{ for some } \|w\| \le B\right\}$$

be a class of linear classifiers on the ball $\mathcal{X}$. Then

$$\mathrm{VC}(\mathcal{H}_B) \le B^2R^2$$

Why are we interested in the VC-dimension at all?

Fact II.10. *Let $\mathcal{H}$ be a family of functions taking values in $\{-1,1\}$ with VC-dimension V. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $h \in \mathcal{H}$.*

$$R(h) \le R_{emp}(h) + \sqrt{\frac{V}{N}(\log N - \log\delta)}.$$

What the fuck does this mean?

Corollary II.11. *For any $h \in \mathcal{H}_B$ defined in fact II.9, with probability at least $1-\delta$,*

$$R(h) \le R_{emp}(h) + O\left(\frac{RB}{\sqrt{N}}\right)$$

Where did the $\log N$ go?

This motivates the following formulation of the SVM problem for linearly non-separable data, since smaller w gives smaller bounds on the error.

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\Big(1 - y_i(\langle w,\, x^{(i)}\rangle + b)\Big)_+$$

where $(x)_+ = x[x \ge 0] = 0 \vee x = \max(0,x)$, and C is a penalty for wrong answers.

II.3 Nonseparable SVM

Lecture 5. Sunday March 02

We can rewrite this more conveniently (without the ugly max function) by introducing *slack variables* $\xi_i \geq 0$ for each $i \in [n]$.

$$\boxed{\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i \quad \text{subject to} \quad \begin{cases} y_i\Big(\langle w,\, x^{(i)}\rangle + b\Big) \geq 1 - \xi_i, \\ \xi_i \geq 0. \end{cases}}$$

Now **SOLVE!**

The Lagrangian is

$$\begin{aligned} L(w, b, \xi, \lambda^{(1)}, \lambda^{(2)}) = {} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i \\ & - \sum_{i=1}^n \lambda_i^{(1)}\Big(y_i(\langle w,\, x^{(i)}\rangle + b) - 1 + \xi_i\Big) - \sum_{i=1}^n \lambda_i^{(2)}\xi_i. \end{aligned}$$

For the KKT point, the stationary conditions are

$$0 = \nabla_w L = w - \sum_{i=1}^n \lambda_i^{(1)} y_i x^{(i)}, \tag{II.4}$$

$$0 = \nabla_b L = -\sum_{i=1}^n \lambda_i^{(1)} y_i, \tag{II.5}$$

$$0 = \nabla_\xi L = C - \lambda^{(1)} - \lambda^{(2)}. \tag{II.6}$$

The complementary slackness conditions are

$$0 = \lambda_i^{(1)}\Big(y_i(\langle w,\, x^{(i)}\rangle + b) - 1 + \xi_i\Big), \tag{II.7}$$

$$0 = \lambda_i^{(2)}\xi_i. \tag{II.8}$$

II.3.1 Wolfe Dual

The Wolfe dual is

$$\max_{w,b,\xi,\lambda^{(1)},\lambda^{(2)}} L(w, b, \xi, \lambda^{(1)}, \lambda^{(2)}) \quad \text{subject to} \quad \begin{cases} \lambda^{(1)} \geq 0, \\ \lambda^{(2)} \geq 0, \\ \lambda_i^{(1)} + \lambda_i^{(2)} = C, \\ w = \sum_{i=1}^n \lambda_i^{(1)} y_i x^{(i)}, \\ \sum_{i=1}^n \lambda_i^{(1)} y_i = 0. \end{cases}$$

Lecture 5: Quadradic programming formulation of soft-margin SVM

Substituting equations (II.5) and (II.6) into L,

$$L^* = -\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \lambda_i^{(1)} \langle w,\, y_i x^{(i)} \rangle + \sum_{i=1}^{n} \lambda_i^{(1)}.$$

Substituting equation (II.4),

$$L^* = \sum_i \lambda_i^{(1)} - \frac{1}{2} \sum_{i,j} \lambda_i^{(1)} \lambda_j^{(1)} y_i y_j \langle x^{(i)},\, x^{(j)} \rangle.$$

We have seen this before! In the linearly separable case, the only difference was that $\lambda_i^{(1)}$ were positive unrestricted, but here they are bounded above by C, because of equation (II.6). Thus the Wolfe dual boils down to

$$\boxed{\max_{0 \le \lambda^{(1)} \le C} \sum_i \lambda_i^{(1)} - \frac{1}{2} \sum_{i,j} \lambda_i^{(1)} \lambda_j^{(1)} y_i y_j \langle x^{(i)},\, x^{(j)} \rangle.}$$

Equation (II.6) is very interesting. For each i, $\lambda_i^{(1)} + \lambda_i^{(2)} = C$.

- If $\lambda_i^{(1)} = 0 \iff \lambda_i^{(2)} = C$, then $\xi_i = 0$ by equation (II.8). This gives $y_i(\langle w,\, x^{(i)} \rangle + b) \ge 1$ for the constraints to hold.
- If $0 < \lambda_i^{(1)} < C \iff 0 < \lambda_i^{(2)} < C$, then $\xi_i = 0$ by equation (II.8). But from equation (II.7), $y_i(\langle w,\, x^{(i)} \rangle + b) = 1$.
- If $\lambda_i^{(1)} = C \iff \lambda_i^{(2)} = 0$, then $0 \le \xi_i = (1 - y_i(\langle w,\, x^{(i)} \rangle + b)) \vee 0$.

Also note that $\xi_i > 0$ is possible only in the last case, $\lambda_i^{(1)} = C$ or $\lambda_i^{(2)} = 0$. This is also the only case where $y_i\big(\langle w,\, x^{(i)} \rangle + b\big) < 1$. This makes sense, because for the objective function to be minimized, ξ_i needs to be as small as possible. There is no need to have a positive ξ_i if the constraints are satisfied without it.