Operations Research III: Theory

Case Study: Regression Models and Support Vector Machine

Ling-Chieh Kung

Department of Information Management
National Taiwan University

Applications of Operations Research theories

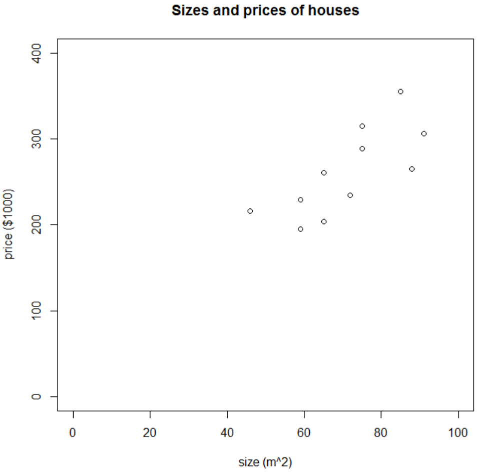
- The theory of Operations Research has been used to develop **models in many fields**.
  - The doctoral degree obtained by George Dantzig (the inventor of the simplex method) is Statistics.
- We will introduce some models in **Statistics** and **Machine Learning** developed with Operations Research.
  - More specifically, Nonlinear Programming.

Road map

- **Regression models**.
- Support Vector Machine.

Linear regression

- Consider a set of data (x_i, y_i), $i = 1, ..., n$.
- If we believe that x_i and y_i has a linear relationship, we may apply **simple linear regression** to fit these data.

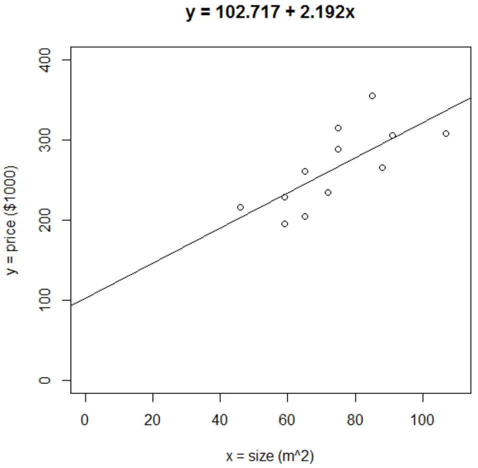


Linear regression

- We try to find α and β such that a line $y = \alpha + \beta x$ to **minimize the sum of squared errors** for all the data points:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \Big[y_i - (\alpha + \beta x_i) \Big]^2.$$

- Note that this is to solve a nonlinear program:
  - Is this a convex program?
  - May we solve it?



Linear regression: convexity

- Let $f(\alpha, \beta) = \sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)]^2$ be the objective function. We have

$$\nabla f = \begin{bmatrix} -2\sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)] \\ -2\sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)]x_i \end{bmatrix} \text{ and } \nabla^2 f = \begin{bmatrix} 2n & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 \end{bmatrix}.$$

- The objective function is convex as $n > 0$ and

$$|\nabla^2 f| = n\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i = (n-1)\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} x_i x_j$$
$$= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i - x_j)^2 \geq 0.$$

Linear regression: convexity

- Alternatively, we may denote

$$f(\alpha, \beta) = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2 = \sum_{i=1}^{n} f_i(\alpha, \beta)$$

  and see that

$$f_i(\alpha, \beta) = y_i^2 + \alpha^2 + \beta^2 x_i^2 - 2\alpha y_i - 2\beta x_i y_i + 2\alpha\beta x_i.$$

- We then have

$$\nabla^2 f_i(\alpha, \beta) = \begin{bmatrix} 2 & 2x_i \\ 2x_i & 2x_i^2 \end{bmatrix},$$

  which means f_i is a convex function.
- As the summation of convex functions is also convex, f is convex.

Linear regression: solution

- Simple linear regression is to solve an unconstrained convex program.
- Numerical algorithms may be used to solve for the optimal α and β.
- Nevertheless, analysis gives us a **closed-form formula**:
  - $\nabla f(\alpha, \beta) = 0$ requires

$$-2\sum_{i=1}^{n}\Big[y_i - (\alpha + \beta x_i)\Big] = 0 \quad \text{and} \quad -2\sum_{i=1}^{n}\Big[y_i - (\alpha + \beta x_i)\Big]x_i = 0$$

    which implies

$$n\alpha + \left(\sum_{i=1}^{n} x_i\right)\beta = \sum_{i=1}^{n} y_i \quad \text{and} \quad \left(\sum_{i=1}^{n} x_i\right)\alpha + \left(\sum_{i=1}^{n} x_i^2\right)\beta = \sum_{i=1}^{n} x_i y_i.$$

  - Solving the linear system results in a direct way to optimize α and β.
- Complete this and compare your result with your Statistics textbook!

Linear regression: remarks

- The same idea applies to **multiple linear regression**: Given a data set $\{x_1^i, x_2^i, ..., x_p^i, y_i)_{i=1,...,n}$, find α, β_1, β_2, ..., and β_p to solve

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \Big[y_i - (\alpha + \sum_{j=1}^{p} \beta_j x_j^i)\Big]^2 = \min_{\alpha,\beta} \sum_{i=1}^{n} \Big[y_i - (\alpha + \beta^{\mathrm{T}} x^i)\Big]^2.$$

- There are many perspectives to consider linear regression:
  - As solving a nonlinear optimization problem.
  - As projecting a vector to a vector space.
  - And others.
- One reason to define fitting error as the sum of **squared** errors rather than the sum of absolute errors is to reduce the **difficulty of optimization**.

Some other regression models

- When one applies linear regression for prediction and hopes to avoid overfitting, one may apply **regularization**. Let $\lambda > 0$ be the given penalty of "using variables," we have:
  - Ridge regression:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \left[y_i - (\alpha + \beta^{\mathrm{T}} x^i)\right]^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

  - LASSO regression:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \left[y_i - (\alpha + \beta^{\mathrm{T}} x^i)\right]^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

- Both the above two models are solving unconstrained convex programs.

Road map

- Regression models.
- **Support vector machine**.

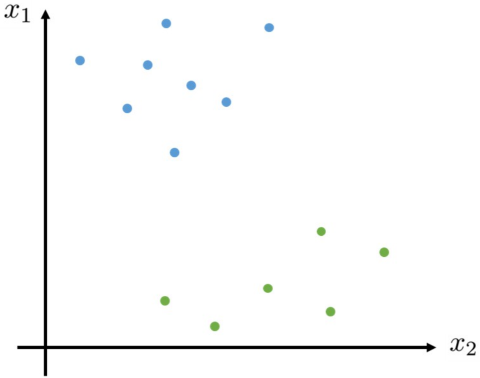
Classification

- **Classification** is an important subject in Machine Learning.
- We are given a data set $\{x_1^i, x_2^i, ..., x_n^i, y_i\}_{i=1,...,m}$, where $y_i \in \{1, -1\}$ labels some kind of success and failure.
- We want to find a **classifier** to assign data point i a **class** according to x^i to minimize the total number of classification error.[1]
- General classification is difficult. Let's do **linear classification**.

[1] It will be better if we state as classification problem as a prediction problem with a slightly different statement. Nevertheless, as this is an optimization course, let's keep things as simple as possible.

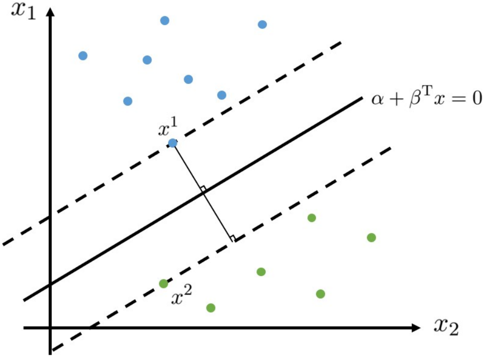
Linear classification

- Let's start with an example.
- Linear classification: How to draw a straight line to **separate** blue dots and green dots?
  - $\mathbb{R}^2$: a line.
  - $\mathbb{R}^3$: a plane.
  - $\mathbb{R}^n$ for $n > 3$: a hyperplane.
- While (infinitely) many lines may do the separation in this example, which one is the "best"?



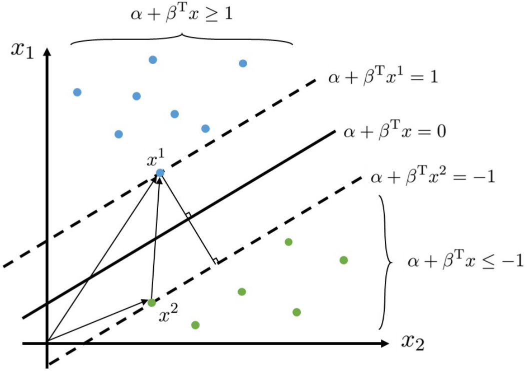
Support vector machine

- The line that is the **farthest** from both groups is the best.
- The two dashed lines are called **supporting hyperplanes**, one for each group.
- The two points x^1 and x^2 (or vectors in general) are called **support vectors**.
- A **support vector machine** (SVM) finds the best separating hyperplane $\alpha + \beta^{\mathrm{T}} x = 0$.



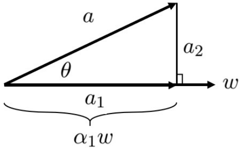
Problem formulation

- Let's classify a point as blue if $\alpha + \beta^{\mathrm{T}} x \geq 1$ or green if $\alpha + \beta^{\mathrm{T}} x \leq -1$.
  - It is equivalent to use k and $-k$ instead of 1 and -1 as we may scale α and β in any way we like.
- The distance between the two supporting hyperplanes is the length of the **projection** of $x^1 - x^2$ onto the normal vector of the separating hyperplane (which is β).



Vector projection

- What is the projection of $a \in \mathbb{R}^n$ onto $w \in \mathbb{R}^n$?
- Let the projection $a_1 = \alpha_1 w$, where $\alpha_1 \in \mathbb{R}$. We then have $\|a_1\| = \|a\| \cos\theta$ and $\|a_1\| = \alpha_1 \|w\|$. They imply that $\alpha_1 = \frac{\|a\| \cos\theta}{\|w\|}$.
  - $\|x\| = \sqrt{x_1^2 + \cdots + x_n^2}$ is the norm (length) of x.



- It then follows that

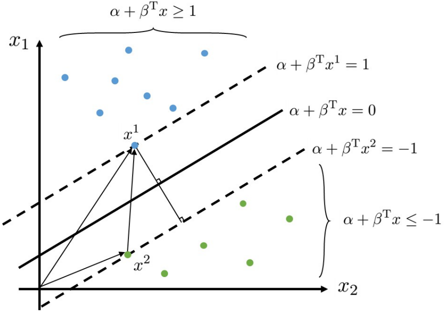
$$a_1 = \alpha_1 w = \frac{\|a\| \cos\theta}{\|w\|} w = \frac{\|a\| \frac{a^{\mathrm{T}} w}{\|a\|\|w\|}}{\|w\|} w = \frac{a^{\mathrm{T}} w}{\|w\|^2} w$$

$$\Rightarrow \|a_1\| = \frac{a^{\mathrm{T}} w}{\|w\|^2} \|w\| = \frac{a^{\mathrm{T}} w}{\|w\|}.$$

Problem formulation

- The length of the projection of a onto w is $\frac{a^{\mathrm{T}}w}{\|w\|}$.
- For $x^1 - x^2$ and β, that length is $\frac{(x^1-x^2)^{\mathrm{T}}\beta}{\|\beta\|}$.
- The objective function is thus

$$\max_{\alpha,\beta} \frac{(x^1-x^2)^{\mathrm{T}}\beta}{\|\beta\|}.$$



- α and β should ensure that, for all $i = 1, ..., m$, we have

$$y_i(\alpha + \beta^{\mathrm{T}}x^i) \geq 1.$$

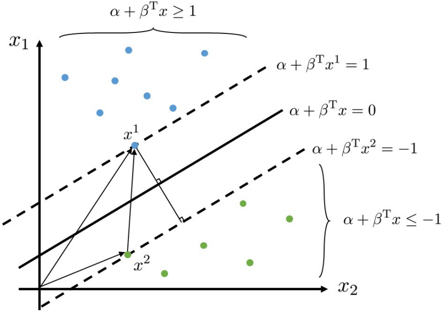
Simplifying the objective function

- The objective function is

$$\max_{\alpha,\beta} \frac{(x^1 - x^2)^{\mathrm{T}}\beta}{\|\beta\|}.$$

- However, given n data points, how may we know which two are supporting?
- Luckily, as x^1 and x^2 are supporting, we have $(x^1 - x^2)^{\mathrm{T}}\beta = 2$.
- The objective function becomes

$$\max_{\alpha,\beta} \frac{2}{\|\beta\|}.$$



The SVM problem (version 1)

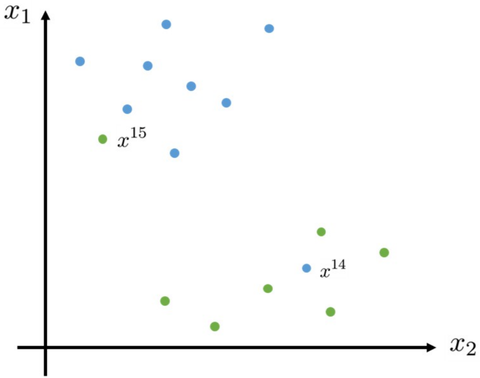
- No one is happy to have decision variables in a denominator.
- Rather than maximizing $\frac{2}{\|\beta\|}$, let's minimize $\frac{1}{2}\|\beta\|$, which is equivalent to minimize $\frac{1}{2}\sum_{k=1}^{n}\beta_k^2$.
- The SVM problem is finally formulated as

$$
\begin{aligned}
\min_{\alpha,\beta} \quad & \frac{1}{2}\sum_{k=1}^{n}\beta_k^2 \\
\text{s.t.} \quad & y_i(\alpha + \beta^{\mathrm{T}} x^i) \geq 1 \qquad \forall i = 1, ..., m.
\end{aligned}
$$

- Note that this is a **convex program**! We may apply numerical algorithms for (constrained) convex programs to solve it.

Imperfect separation

- In many case **perfect separation** (with no classification error) is impossible.
- In this case, we allow errors but add the "degree of errors" into the objective function.



The SVM problem (final version)

- Given a separating hyperplane $\alpha + \beta^{\mathrm{T}} x = 0$, ideally we have $y_i(\alpha + \beta^{\mathrm{T}} x^i) \geq 1$ for data point i.
- When this is violated, let $\gamma_i \geq 0$ be the **degree of violation**.
- The SVM problem that allows imperfect separation is

$$\begin{aligned} \min_{\alpha, \beta, \gamma} \quad & \frac{1}{2} \sum_{k=1}^{n} \beta_k^2 + C \sum_{i=1}^{m} \gamma_i \\ \text{s.t.} \quad & y_i(\alpha + \beta^{\mathrm{T}} x^i) \geq 1 - \gamma_i \quad \forall i = 1, ..., m \\ & \gamma_i \geq 0 \quad \forall i = 1, ..., m, \end{aligned}$$

  where $C \geq 0$ is a given parameter.
  - A larger C means a larger **penalty** is incurred with classification errors.
- This is still a **convex program**!

Dualization for the SVM problem

- To solve this constrained convex program, let's find its **Lagrange dual program**.
- Let $\lambda_i \geq 0$ and $\mu_i \geq 0$ be the Lagrange multipliers, the Lagrangian is

$$\begin{aligned}\mathcal{L}(\alpha, \beta, \gamma | \lambda, \mu) = {} & \frac{1}{2}\sum_{k=1}^{n} \beta_k^2 + C\sum_{i=1}^{m} \gamma_i \\ & - \sum_{i=1}^{m} \lambda_i \left[y_i \left(\alpha + \sum_{k=1}^{n} x_k^i \beta_k \right) - 1 + \gamma_i \right] - \sum_{i=1}^{m} \mu_i \gamma_i.\end{aligned}$$

- The Lagrange dual program is

$$\max_{\lambda \geq 0, \mu \geq 0} \min_{\alpha, \beta, \gamma} \mathcal{L}(\alpha, \beta, \gamma | \lambda, \mu).$$

Analyzing the inner program

- To choose α, β_k, and γ_i to minimize

$$\frac{1}{2}\sum_{k=1}^{n}\beta_k^2 + C\sum_{i=1}^{m}\gamma_i - \sum_{i=1}^{m}\lambda_i\left[y_i\left(\alpha + \sum_{k=1}^{n}x_k^i\beta_k\right) - 1 + \gamma_i\right] - \sum_{i=1}^{m}\mu_i\gamma_i,$$

  the first-order condition is necessary and sufficient:

$$\sum_{i=1}^{m}\lambda_i y_i = 0, \quad \beta_k = \sum_{i=1}^{m}\lambda_i y_i x_k^i \ \forall k, \text{ and } \quad C = \lambda_i + \mu_i \ \forall i.$$

- The first and third sets of constraints do not have any primal variable involved. They will become **constraints** of the Lagrangian dual program for the **dual variables** to satisfy.

Analyzing the inner program

- For any $\lambda \geq 0$ and $\mu \geq 0$ satisfying $\sum_{i=1}^{m} \lambda_i y_i = 0$ and $C = \lambda_i + \mu_i$ for all $i = 1, ..., m$, the Lagrangian can be simplified:

$$\frac{1}{2}\sum_{k=1}^{n}\beta_k^2 + C\sum_{i=1}^{m}\gamma_i - \sum_{i=1}^{m}\lambda_i\left[y_i\left(\alpha + \sum_{k=1}^{n}x_k^i\beta_k\right) - 1 + \gamma_i\right] - \sum_{i=1}^{m}\mu_i\gamma_i$$

$$= \frac{1}{2}\sum_{k=1}^{n}\beta_k^2 - \sum_{i=1}^{m}\lambda_i\left[y_i\sum_{k=1}^{n}x_k^i\beta_k\right] + \sum_{i=1}^{m}\lambda_i.$$

- Plugging $\beta_k = \sum_{j=1}^{m} \lambda_j y_j x_k^j$ into the Lagrangian results in

$$\frac{1}{2}\sum_{k=1}^{n}\left(\sum_{j=1}^{m}\lambda_j y_j x_k^j\right)^2 - \sum_{i=1}^{m}\lambda_i y_i\sum_{k=1}^{n}\left(\sum_{j=1}^{m}\lambda_j y_j x_k^j\right)x_k^i + \sum_{i=1}^{m}\lambda_i$$

$$= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j y_i y_j (x^i)^{\mathrm{T}}x^j + \sum_{i=1}^{m}\lambda_i.$$

Dualization for the SVM problem

- The Lagrangian dual program now becomes

$$\begin{aligned} \max_{\lambda,\mu} \quad & -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j y_i y_j (x^i)^{\mathrm{T}} x^j + \sum_{i=1}^{m}\lambda_i \\ \text{s.t.} \quad & \sum_{i=1}^{m}\lambda_i y_i = 0 \\ & C = \lambda_i + \mu_i \quad \forall i = 1, ..., m \\ & \lambda_i \geq 0,\ \mu_i \geq 0 \quad \forall i = 1, ..., m. \end{aligned}$$

- To further simplify this program, note that μ_i does not exist in the objective function.
- In fact, it simply tells us that λ_i cannot be greater than C.

The dual program of the SVM problem

- The dual program of is finally derived as

$$\begin{aligned} \max_{\lambda} \quad & -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j y_i y_j (x^i)^{\mathrm{T}} x^j + \sum_{i=1}^{m}\lambda_i \\ \text{s.t.} \quad & \sum_{i=1}^{m}\lambda_i y_i = 0 \\ & 0 \leq \lambda_i \leq C \quad \forall i = 1, ..., m. \end{aligned}$$

- This is another constrained nonlinear program.
  - The number of variables and constraints are m and $1 + 2m$. Those for the primal are $1 + n + m$ and $2m$.
  - Most of the dual constraints are "simple".
  - Nevertheless, is it really a convex program (those the theory tells us so)?

Convexity of the dual program

- Let

$$f(\lambda) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j y_i y_j (x^i)^{\mathrm{T}} x^j - \sum_{i=1}^{m}\lambda_i$$

be the negation of the objective function, we have

$$\begin{aligned}\nabla^2 f(\lambda) &= \begin{bmatrix} y_1y_1(x^1)^{\mathrm{T}}x^1 & y_1y_2(x^1)^{\mathrm{T}}x^2 & \cdots & y_1y_m(x^1)^{\mathrm{T}}x^m \\ y_2y_1(x^2)^{\mathrm{T}}x^1 & & & \\ \vdots & & \ddots & \vdots \\ y_my_1(x^m)^{\mathrm{T}}x^1 & \cdots & & y_my_m(x^m)^{\mathrm{T}}x^m \end{bmatrix} \\ &= \begin{bmatrix} y_1x_1^{\mathrm{T}} \\ \vdots \\ y_mx_m^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} y_1x_1 & \cdots & y_mx_m \end{bmatrix} = Z^{\mathrm{T}}Z,\end{aligned}$$

where $Z = \begin{bmatrix} y_1x_1 & \cdots & y_mx_m \end{bmatrix} \in \mathbb{R}^{n\times m}$.

Convexity of the dual program

- To show that the Hessian is positive semidefinite, we use the definition. Because

$$x^{\mathrm{T}}\nabla^2 f(\lambda)x = x^{\mathrm{T}}Z^{\mathrm{T}}Zx = (Zx)^{\mathrm{T}}Zx = \|Zx\|^2 \geq 0 \quad \forall x \in \mathbb{R}^m,$$

  the proof is complete.

Remarks

- In this lecture, we show how the theory of Operations Research may be utilized to develop **models** in related fields.
  - There are much more!
  - We choose the examples from Statistics and Machine Learning not because these are the most important.
  - We do so because at this moment many students want to learn these subjects.
  - A **solid foundation** is needed for us to get deeper understanding.
- There are still a lot of interesting things to learn.
  - Now you have a not-so-bad foundation.
  - Go and explore the fascinating world!

That's all. Thank you!