Foundations of Data Science, Fall 2020

4. Maximum Likelihood

Prof. Dan Olteanu

DaST Data•(Systems+Theory)

University of Zurich[UZH]

Sept 29, 2020

https://lms.uzh.ch/url/RepositoryEntry/16830890400

https://uzh.zoom.us/j/96690150974?pwd=cnZmMTduWUtCeWoxYW85Z3RMYnpTZz09

Recap from Last Lecture: Predicting Commute Time

Goal

- Predict the time taken for commute given distance and day of week
- Do we only wish to make predictions or also suggestions?

Model and Choice of Loss Function

- Use a linear model

$$y = w_0 + w_1 x_1 + \cdots + w_D x_D + \epsilon = \widehat{y} + \epsilon$$

- Minimise average squared error $\frac{1}{2N} \sum (y_i - \widehat{y}_i)^2$

Algorithm to Fit Model

- Simple matrix operations using closed-form solution

What We Left Out in the Previous Lecture

- How to model the noise ϵ in the model
- Beyond linear models: Non-linearity using basis expansion
- What to do when you have more features than data?

We will treat these aspects in the upcoming lectures.

Model and Loss Function Choice

"Optimisation" View of Machine Learning

- Pick model that you expect may fit the data well enough
- Pick a measure of performance that makes "sense" and can be optimised
- Run optimisation algorithm to obtain model parameters

Probabilistic View of Machine Learning

- Pick a model for data and explicitly formulate the deviation (or uncertainty) from the model using the language of probability
- Use notions from probability to define suitability of various models
- "Find" the parameters or make predictions on unseen data using these suitability criteria

Probabilistic Perspective of Linear Regression: Road Map

- Maximum Likelihood Principle
- Probabilistic Formulation of the Linear Model via Maximum Likelihood
- Maximum Likelihood and Least Squares Estimate

Maximum Likelihood Principle

Given: Observations $x_1, \ldots, x_N$ drawn independently from the same distribution p

- i.i.d. = independently and identically distributed
- p has a parametric form, e.g., the parameter vector $\boldsymbol{\theta}$

Likelihood of observing $x_1, \ldots, x_N$ = probability of making these observations assuming they are generated according to p

- $p(x_1, \ldots, x_N | \boldsymbol{\theta})$: joint probability distribution of the observations given $\boldsymbol{\theta}$

Maximum Likelihood Principle: Pick parameter $\boldsymbol{\theta}$ that maximises the likelihood

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta}} p(x_1, \ldots, x_N \mid \boldsymbol{\theta})$$

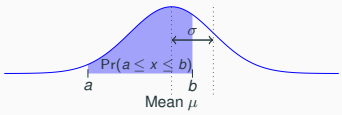
Since the observations are i.i.d.: $p(x_1, \ldots, x_N \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p(x_i \mid \boldsymbol{\theta})$

Univariate Gaussian (Normal) Distribution

The univariate normal distribution is defined by the following density function

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad X \sim \mathcal{N}(\mu, \sigma^2)$$

Mean μ, standard deviation σ, variance σ^2



$$\int_{-\infty}^{\infty} p(x)\,dx = 1$$
$$\int_{-\infty}^{\infty} xp(x)\,dx = \mu$$
$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,dx = \sigma^2$$

Standardisation: By setting $Y = \frac{X-\mu}{\sigma}$, $Y \sim \mathcal{N}(0,1)$

Play: NormalIntro.xls on OLAT (from Khan Academy)

Covariance

For random variables X and Y the covariance measures how they change jointly

$$\mathrm{cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])\right]$$

Suppose $\mathbf{x}$ is a D-dimensional random vector. The covariance matrix consists of all pairwise covariances.

$$\mathrm{cov}(\mathbf{x}) = \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\mathsf{T}\right] = \begin{bmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_D) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2, X_D) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_D, X_1) & \mathrm{cov}(X_D, X_2) & \cdots & \mathrm{var}(X_D) \end{bmatrix}.$$

Multivariate Normal (Gaussian) Distribution

Suppose $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $1 \le i \le N$ are independent

The joint probability distribution $p(x_1, \ldots, x_N)$ is a multivariate normal distribution

$$p(x_1, \ldots x_N) = \prod_{i=1}^{N} p(x_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \cdot \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$$

$$= \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \cdot \exp\left(-\frac{1}{2} \cdot (\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

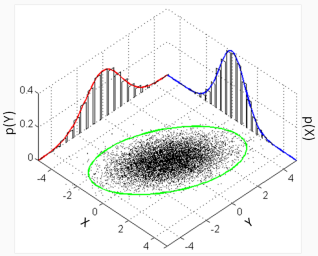
where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_N^2 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

Bivariate Normal (Gaussian) Distribution

$$p(X, Y) = p(X) \cdot p(Y) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} \cdot \exp\left(-\frac{1}{2} \cdot (\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} X \\ Y \end{bmatrix}$$



All equiprobable points lie on an ellipse.

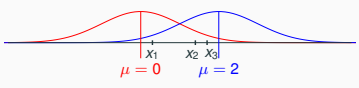
Maximum Likelihood Example 1

- Given: Three independent samples $x_1 = 0.3$, $x_2 = 1.4$, and $x_3 = 1.7$

  Likelihood: $p(x_1, x_2, x_3 \mid \boldsymbol{\theta}) = p(x_1 \mid \boldsymbol{\theta}) \cdot p(x_2 \mid \boldsymbol{\theta}) \cdot p(x_3 \mid \boldsymbol{\theta})$

- Data generated from normal distributions: either $\mathcal{N}(0,1)$ or $\mathcal{N}(2,1)$

$$\boldsymbol{\theta} = [\mu, \sigma] = \begin{cases} [0,1]^\mathsf{T} \\ [2,1]^\mathsf{T} \end{cases}$$



Which distribution is more likely?

Maximum Likelihood: Example 2

Discrete random variable X with the following probability distribution ($\theta \in (0,1)$)

X	0	1	2	3
$p(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

Observations: (3,0,2,1,3,2,1,0,2,1).

Compute the maximum likelihood estimate for θ.

$$L(x_1, \ldots, x_{10} \mid \theta) = \prod_{i=1}^{10} P(x_i \mid \theta)$$

$$= P(X=3)^2 \cdot P(X=0)^2 \cdot P(X=2)^3 \cdot P(X=1)^3$$

$$= \left(\frac{1-\theta}{3}\right)^2 \cdot \left(\frac{2\theta}{3}\right)^2 \cdot \left(\frac{2(1-\theta)}{3}\right)^3 \cdot \left(\frac{\theta}{3}\right)^3,$$

$$\log L = 2 \cdot \left(\log(1-\theta) - \log 3\right) + 2\left(\log 2\theta - \log 3\right) + 3\left(\log 2(1-\theta) - \log 3\right) + 3 \cdot \left(\log\theta - \log 3\right)$$

$\log L(\theta) = \text{Constant} + \sum \log\theta + \sum \log(1-\theta)$

$\frac{\partial \log L}{\partial \theta} = \frac{\sum}{\theta} + \frac{\sum}{1-\theta}\cdot(-1) = 0 \Rightarrow \frac{\sum}{\theta} = \frac{\sum}{1-\theta}$

$\Rightarrow \theta = 1-\theta \Rightarrow \boxed{\theta = 1/2}$

Maximum Likelihood: Example 3

Assume N heights $x_1, \ldots, x_N$ are i.i.d. random variables drawn from $\theta = [\mu, \sigma]^T$.

What is the maximum likelihood estimator of $p(x_1, \ldots, x_N \mid \theta)$?

$p(x_i \mid \theta) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$

$p(x_1, \ldots, x_N \mid \theta) = \prod_{i\in[N]} p(x_i \mid \theta) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^N \cdot \exp\left(-\frac{1}{2\sigma^2}\cdot\sum_i (x_i-\mu)^2\right)$

$\mathcal{L}(\theta) = \ln p(x_1, \ldots, x_N \mid \theta) = -N\cdot(\ln\sqrt{2\pi} + \ln\sigma) - \frac{1}{2\sigma^2}\cdot\sum_i (x_i-\mu)^2$

LL

$\frac{\partial \mathcal{L}}{\partial \mu} = -\frac{1}{2\sigma^2}\cdot\sum_i (x_i-\mu)\cdot(-1) = \frac{1}{\sigma^2}\cdot\sum_i (x_i-\mu)$

$= \frac{1}{\sigma^2}\cdot\left(\sum_i x_i - \sum_i \mu\right)$ ($\sum_i \mu = N\cdot\mu$)

$\frac{\partial \mathcal{L}}{\partial \mu} = 0 \Rightarrow \frac{1}{\sigma^2}\cdot\left(\sum_i x_i - N\cdot\mu_{ML}\right) = 0 \Rightarrow \boxed{\mu_{ML} = \frac{\sum_i x_i}{N}}$

$\mathcal{L}(\theta) = -N\left(\ln\sqrt{2\pi} + \ln\sigma\right) - \frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2$

$\frac{\partial \mathcal{L}}{\partial \sigma} = -\frac{N}{\sigma} - (-2)\cdot\frac{1}{2}\cdot\frac{1}{\sigma^3}\cdot\sum_i (x_i-\mu)^2$

$\frac{\partial \mathcal{L}}{\partial \sigma} = 0 \Rightarrow -\frac{N}{\sigma} + \frac{1}{\sigma^3}\cdot\left(\sum_i x_i - \mu_{ML}\right)^2 = 0$

$\boxed{\sigma^2_{ML} = \frac{\sum_i (x_i-\mu_{ML})^2}{N}}$

Linear Regression

Linear Model

$$y = w_0 x_0 + w_1 x_1 + \cdots + w_D x_D + \epsilon = \mathbf{w}\cdot\mathbf{x} + \epsilon$$

Model y given $\mathbf{x}$, $\mathbf{w}$ as a random variable with mean $\mathbf{w}^T\mathbf{x}$.

$$\mathbb{E}[y \mid \mathbf{x}, \mathbf{w}] = \mathbf{w}^T\mathbf{x}$$

We will be specific in choosing the distribution of y given $\mathbf{x}$ and $\mathbf{w}$.
Assumption: Given $\mathbf{x}$ and $\mathbf{w}$, y is normal with mean $\mathbf{w}^T\mathbf{x}$ and variance σ^2

$$p(y \mid \mathbf{w}, \mathbf{x}) = \mathcal{N}(\mathbf{w}^T\mathbf{x}, \sigma^2) = \mathbf{w}^T\mathbf{x} + \mathcal{N}(0, \sigma^2)$$

Alternatively, we may view this model as $\epsilon \sim \mathcal{N}(0, \sigma^2)$ (Gaussian Noise)

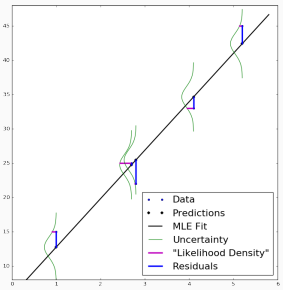
Discriminative Framework

The input $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ is fixed, we do not model a distribution over $\mathbf{X}$

Likelihood of Linear Regression (Gaussian Noise Model)

Observed data $(\mathbf{X}, \mathbf{y})$ made up of $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$

What is the likelihood of observing the data for model parameters $\mathbf{w}$, σ?



MLE Estimator

Find parameters which maximise the likelihood.

(product of "likelihood density" — segments)

Least Square Estimator

Find parameters which minimise the sum of squares of the residuals

(sum of squares of the | segments).

Likelihood of Linear Regression (Gaussian Noise Model)

Observed data $(\mathbf{X}, \mathbf{y})$ made up of $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$

What is the likelihood of observing the data for model parameters $\mathbf{w}$, σ?

$$\boxed{p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = p\left(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w}, \sigma\right) = \prod_{i=1}^N p\left(y_i \mid \mathbf{x}_i, \mathbf{w}, \sigma\right)}$$

According to the model $y_i \sim \mathcal{N}(\mathbf{w}^T\mathbf{x}_i, \sigma^2)$

$$p\left(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w}, \sigma\right) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{N/2}\exp\left(-\frac{1}{2\sigma^2}\cdot\underbrace{\sum_{i=1}^N (y_i - \mathbf{w}^T\mathbf{x}_i)^2}_{(\mathbf{X}\mathbf{w}-\mathbf{y})^T(\mathbf{X}\mathbf{w}-\mathbf{y})}\right)$$

Likelihood of Linear Regression (Gaussian Noise Model)

Observed data $(\mathbf{X}, \mathbf{y})$ made up of $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$

What is the likelihood of observing the data for model parameters $\mathbf{w}$, σ?

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = p\left(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w}, \sigma\right) = \prod_{i=1}^{N} p\left(y_i \mid \mathbf{x}_i, \mathbf{w}, \sigma\right)$$

According to the model $y_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma^2)$

$$p\left(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w}, \sigma\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2} \cdot \underbrace{\sum_{i=1}^{N} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2}_{(\mathbf{X}\mathbf{w}-\mathbf{y})^\top (\mathbf{X}\mathbf{w}-\mathbf{y})}\right)$$

Goal: Find parameters $\mathbf{w}$ and σ that maximise the likelihood

Likelihood of Linear Regression (Gaussian Noise Model)

Likelihood

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w} - \mathbf{y})^\top (\mathbf{X}\mathbf{w} - \mathbf{y})\right)$$

Maximise Likelihood = Maximise Log-Likelihood ($\log : \mathbb{R}^+ \to \mathbb{R}$ is increasing)

$$\mathrm{LL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w} - \mathbf{y})^\top (\mathbf{X}\mathbf{w} - \mathbf{y})$$

Maximise Log-Likelihood = Minimise Negative Log-Likelihood

$$\mathrm{NLL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w} - \mathbf{y})^\top (\mathbf{X}\mathbf{w} - \mathbf{y})$$

Maximum Likelihood and Least Squares Estimates

We first find $\mathbf{w}$ that minimises the negative log-likelihood

$$\mathrm{NLL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w} - \mathbf{y})^\top (\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{N}{2}\log(2\pi\sigma^2)$$

Recall the objective function we used for the least squares estimate in the previous lecture

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2N}(\mathbf{X}\mathbf{w} - \mathbf{y})^\top (\mathbf{X}\mathbf{w} - \mathbf{y})$$

For minimising with respect to $\mathbf{w}$, the two objectives are the same up to a constant additive and multiplicative factor!

The solution $\mathbf{w}_{\mathrm{ML}}$ to find the maximum likelihood estimator **is the same** as the least squares estimator:

$$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

Maximum Likelihood and Least Squares Estimates

The MLE for σ is given by

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}(\mathbf{X}\mathbf{w}_{\mathrm{ML}} - \mathbf{y})^\top (\mathbf{X}\mathbf{w}_{\mathrm{ML}} - \mathbf{y})$$

Prediction using the MLE for Linear Regression

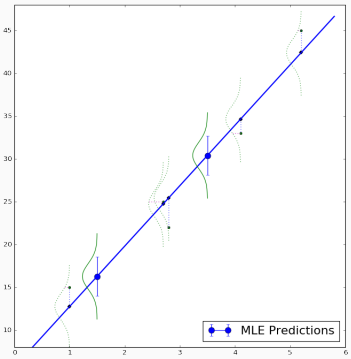
Given training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$, we can obtain the MLE $\mathbf{w}_{\mathrm{ML}}$ and σ_{ML}.

On a new point $\mathbf{x}_{\mathrm{new}}$:

- Make prediction
- Give confidence intervals

$\widehat{y}_{\mathrm{new}} = \mathbf{w}_{\mathrm{ML}} \cdot \mathbf{x}_{\mathrm{new}}$

$y_{\mathrm{new}} \sim \widehat{y}_{\mathrm{new}} + \mathcal{N}(0, \sigma^2_{\mathrm{ML}})$



Summary : MLE for Linear Regression (Gaussian Noise)

Model

- Linear model: $y = \mathbf{w} \cdot \mathbf{x} + \epsilon$
- Explicitly model $\epsilon \sim \mathcal{N}(0, \sigma^2)$

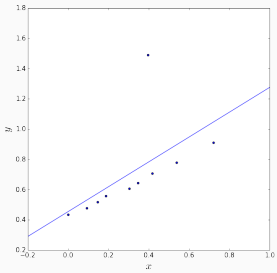
Maximum Likelihood Estimation

- Every $(\mathbf{w}, \sigma)$ defines a probability distribution over observed data
- Pick $\mathbf{w}$ and σ that maximise the likelihood of observing the data

Algorithm

- As in the previous lecture, we have closed form expressions
- Algorithm simply implements elementary matrix operations

Outliers and Laplace Distribution



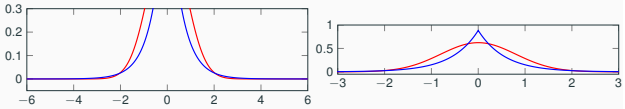
If the data has outliers, we can model the noise using a distribution that has heavier tails

For the linear model $y = \mathbf{w} \cdot \mathbf{x} + \epsilon$, use

$$\epsilon \sim \mathrm{Lap}(0, b),$$

where the density function for $\mathrm{Lap}(\mu, b)$ is given by

$$p(x) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$



Laplace and normal distributions with the same mean and variance

Maximum Likelihood for Laplace Noise Model

Likelihood of observing the data in terms of model parameters $\mathbf{w}$ and b.

$$\begin{aligned} p(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w}, b) &= \prod_{i=1}^{N} \frac{1}{2b} \exp\left(-\frac{|y_i - \mathbf{w}^\mathsf{T}\mathbf{x}_i|}{b}\right) \\ &= \frac{1}{(2b)^N} \exp\left(-\frac{1}{b}\sum_{i=1}^{N} |y_i - \mathbf{w}^\mathsf{T}\mathbf{x}_i|\right) \end{aligned}$$

As for the Gaussian noise model, we look at the negative log-likelihood:

$$\mathrm{NLL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, b) = \frac{1}{b}\sum_{i=1}^{N} |y_i - \mathbf{w}^\mathsf{T}\mathbf{x}_i| + N \log(2b)$$

Maximum likelihood estimate can be obtained by minimising the sum of the absolute values of the residuals, which is the same objective we discussed in the last lecture in the context of fitting a linear model that is robust to outliers.

Maximum Likelihood for Laplace Noise Model

Question: What are $\mathbf{w}_{\mathrm{ML}}$ and b_{ML} that minimise $\mathrm{NLL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, b)$?

- No closed form solution :(
- Also, non-linear objective function, so hard to optimise.

Good news: It can be transformed into a linear objective subject to linear constraints. It is a convex optimisation problem, it has a unique solution.

See page 226 in Murphy book.