Machine Learning - MT 2016
3. Maximum Likelihood

Varun Kanade

University of Oxford
October 17, 2016

Outline

Probabilistic Perspective of Machine Learning

- Probabilistic Formulation of the Linear Model
- Maximum Likelihood Estimate
- Relation to the Least Squares Estimate

Outline

Probability Review

Linear Regression and Maximum Likelihood

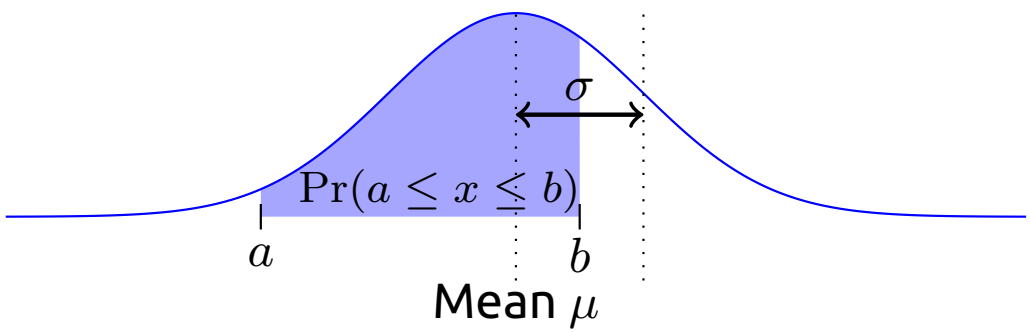
Information, Entropy, KL Divergence

Univariate Gaussian (Normal) Distribution

The univariate normal distribution is defined by the following density function

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

Here μ is the mean and σ^2 is the variance.



$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

$$\int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x = \mu$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,\mathrm{d}x = \sigma^2$$

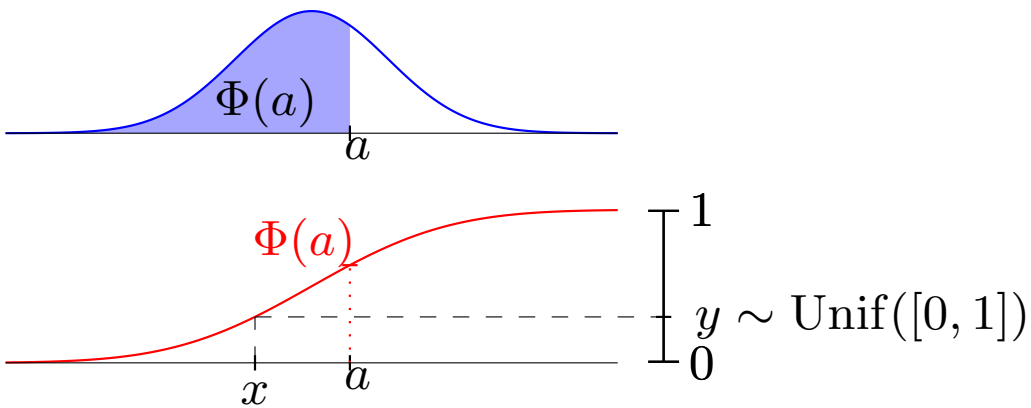
Sampling from a Gaussian distribution

Sampling from $X \sim \mathcal{N}(\mu, \sigma^2)$

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

Cumulative distribution function

$$\Phi(x; 0, 1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$



Bivariate Normal (Gaussian) Distribution

Suppose $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent

The joint probability distribution $p(x_1, x_2)$ is a bivariate normal distribution.

$$
\begin{aligned}
p(x_1, x_2) &= p(x_1) \cdot p(x_2) \\
&= \frac{1}{\sqrt{2\pi}\sigma_1} \cdot \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \cdot \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right) \\
&= \frac{1}{2\pi(\sigma_1^2\sigma_2^2)^{1/2}} \cdot \exp\left(-\left(\frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}\right)\right) \\
&= \frac{1}{2\pi|\mathbf{\Sigma}|^{1/2}} \cdot \exp\left(-\frac{1}{2} \cdot (\mathbf{x}-\boldsymbol{\mu})^\mathsf{T} \mathbf{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)
\end{aligned}
$$

where

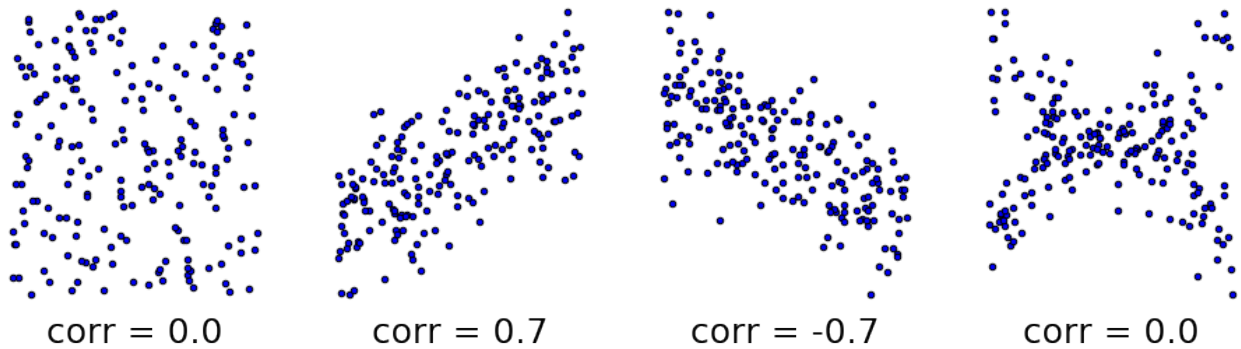
$$
\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}
$$

Note: All equiprobable points lie on an ellipse.

Covariance and Correlation

For random variable X and Y the covariance measures how the random variable change jointly

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



Covariance depends on the scale. The (Pearson) correlation coefficient normalizes the covariance to give a value between -1 and $+1$.

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}},$$

where $\operatorname{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$ and $\operatorname{var}(Y) = \mathbb{E}[(Y - \mathbb{E}[Y])^2]$.

Independent variables are uncorrelated, but the converse is not true!

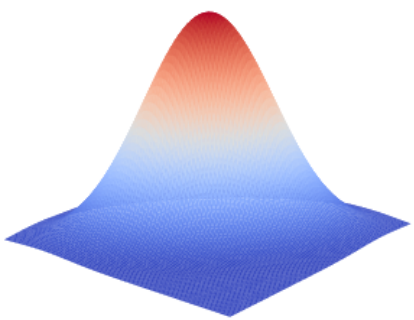
Multivariate Gaussian Distribution

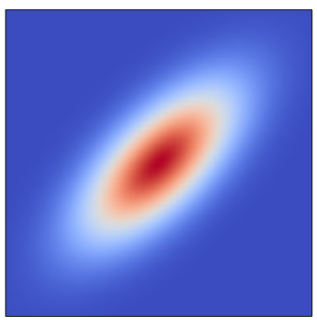
Suppose $\mathbf{x}$ is a D-dimensional random vector. The covariance matrix consists of all pairwise covariances.

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

If $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$ and $\boldsymbol{\Sigma} = \operatorname{cov}(\mathbf{x})$, the multivariate normal is defined by the density

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





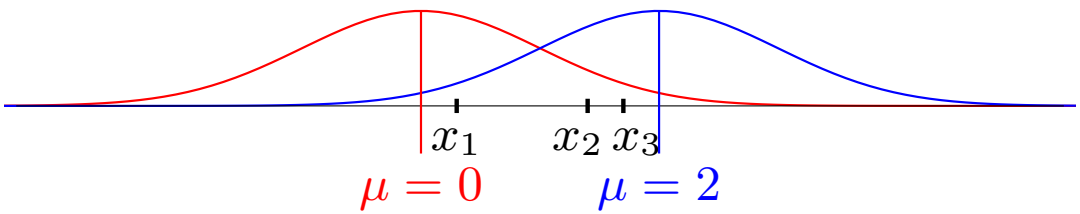
Suppose you are given three independent samples: $x_1 = 0.3$, $x_2 = 1.4$, and $x_3 = 1.7$

You know that the data is generated from either $\mathcal{N}(0, 1)$ or $\mathcal{N}(2, 1)$.

Let $\boldsymbol{\theta}$ represent the parameters (μ, σ) of the two distributions. Then the probability of observing the data with parameter θ is called the likelihood.

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

We have to choose between $\theta = (0, 1)$ or $\theta = (2, 1)$. Which one is more likely?



Maximum Likelihood Estimation (MLE)

Pick parameter θ that maximises the likelihood

Outline

Probability Review

Linear Regression and Maximum Likelihood

Information, Entropy, KL Divergence

Linear Regression

Linear Model

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

Noise/uncertainty

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

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

We will be specific in choosing the distribution of y given $\mathbf{x}$ and $\mathbf{w}$. Let us assume that given $\mathbf{x}, \mathbf{w}$, y is normal with mean $\mathbf{w}^\mathsf{T}\mathbf{x}$ and variance σ^2

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

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

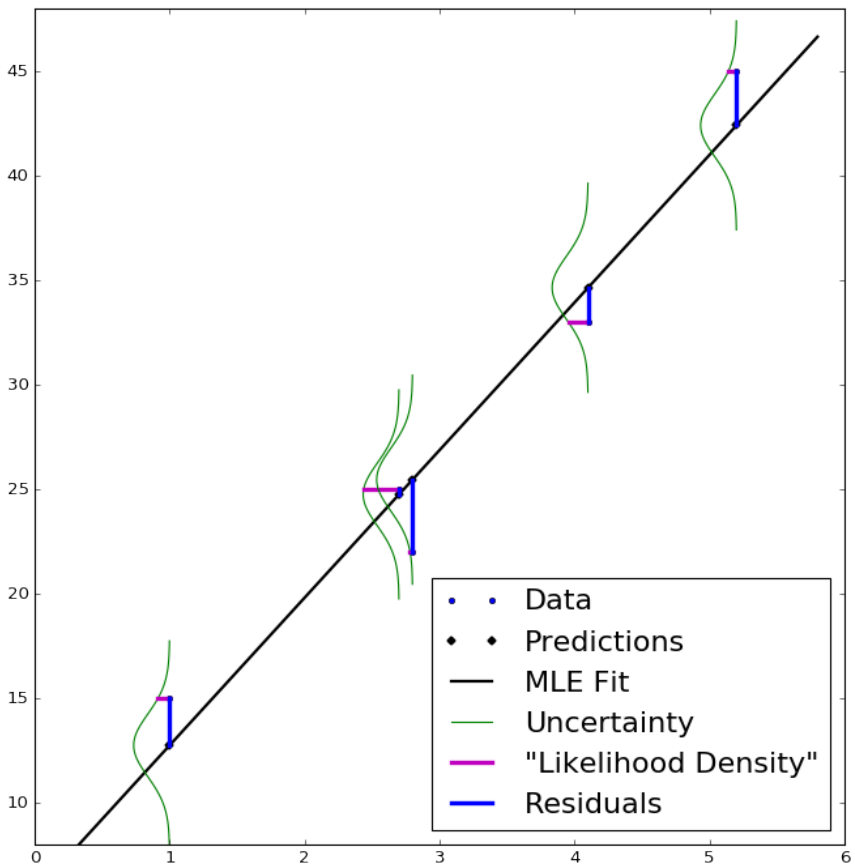
Discriminative Framework

Throughout this lecture, think of the inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N$ as fixed

Likelihood of Linear Regression (Gaussian Noise Model)

Suppose we observe data $\langle (\mathbf{x}_i, y_i) \rangle_{i=1}^{N}$.

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)

Suppose we observe data $\langle (\mathbf{x}_i, y_i) \rangle_{i=1}^N$.

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

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

According to the model $y_i \sim \mathbf{w}^\mathsf{T}\mathbf{x}_i + \mathcal{N}(0, \sigma^2)$

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

Want to find parameters $\mathbf{w}$ and σ that maximise the likelihood

Likelihood of Linear Regression (Gaussian Noise Model)

Let us consider the likelihood $p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma)$

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

As $\log : \mathbb{R}^+ \to \mathbb{R}$ is an increasing function, we can instead maximise the log of the likelihood (called log-likelihood), which results in a simpler mathematical expression.

$$\mathrm{LL}(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w}, \sigma) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{w}^\mathsf{T}\mathbf{x}_i)^2$$

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

Let's first find $\mathbf{w}$ that maximizes the log-likelihood

Maximum Likelihood and Least Squares Estimates

We'd like to find $\mathbf{w}$ that maximises the log-likelihood

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

Alternatively, we can minimise the negative log-likelihood

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

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

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

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

Maximum Likelihood Estimate for Linear Regression

As the solution $\mathbf{w}_{\text{ML}}$ to find the maximum likelihood estimator is the same as the least squares estimator, we have

$$\mathbf{w}_{\text{ML}} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$$

Recall the form of the negative log-likelihood

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

We can also find the maximum likelihood estimate for σ

Exercise on sheet 2 to show that the MLE of σ is given by

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

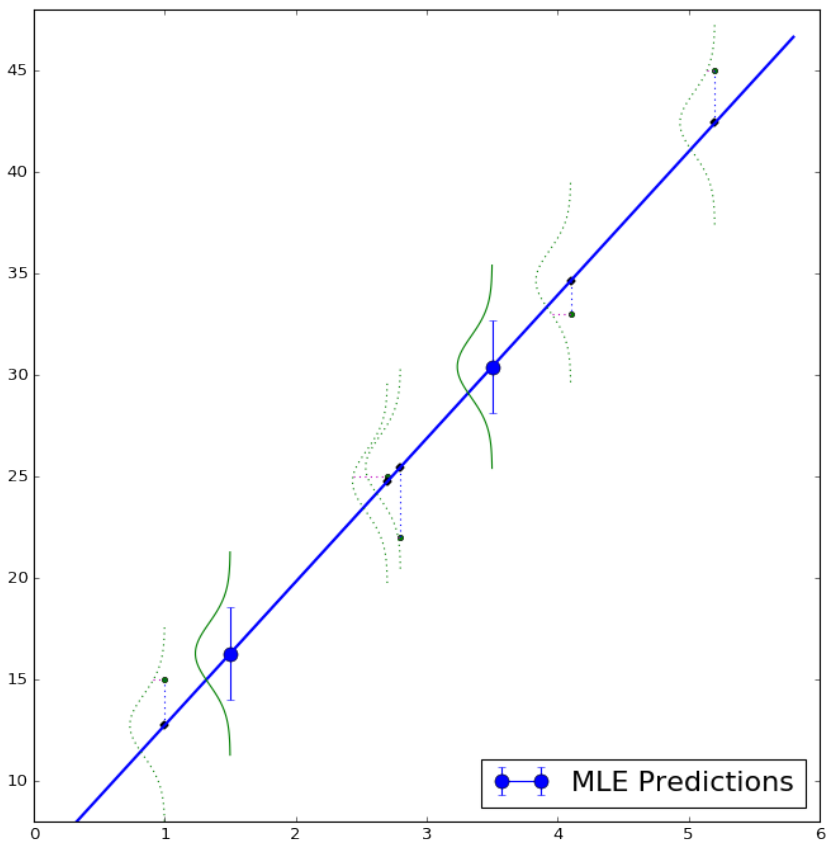
Prediction using the MLE for Linear Regression

Given training data $\langle (\mathbf{x}_i, y_i) \rangle_{i=1}^N$, we can obtain the MLE $\mathbf{w}_{\text{ML}}$ and σ_{ML}.

One a new point $\mathbf{x}_{\text{new}}$, we can use these to make a prediction and also give confidence intervals

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

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



Summary : MLE for Linear Regression (Gaussian Noise)

Model

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

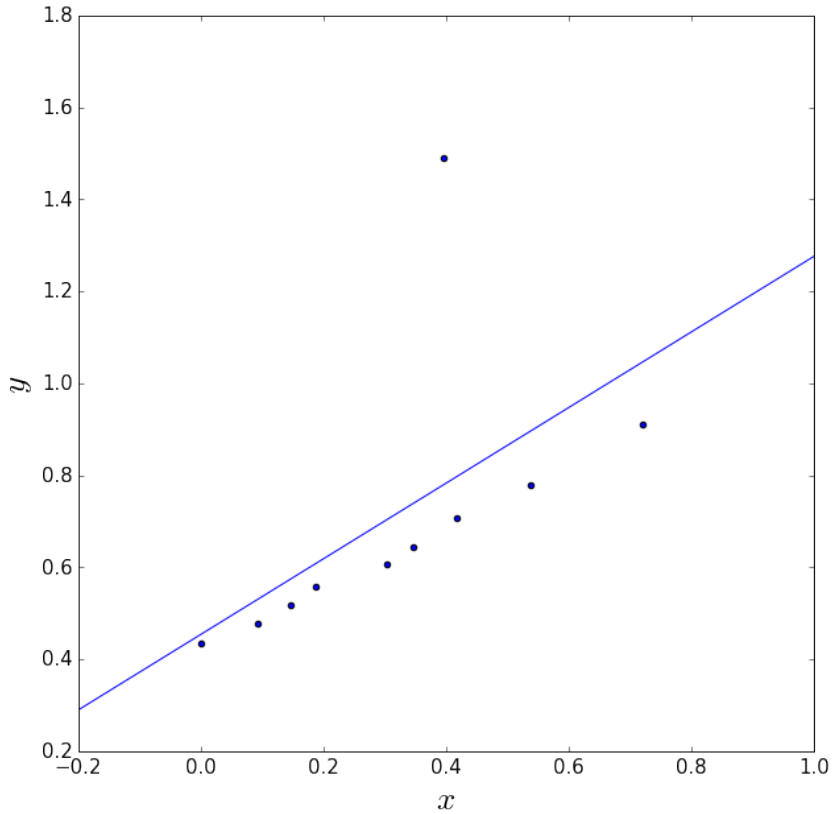
Maximum Likelihood Estimation

- 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



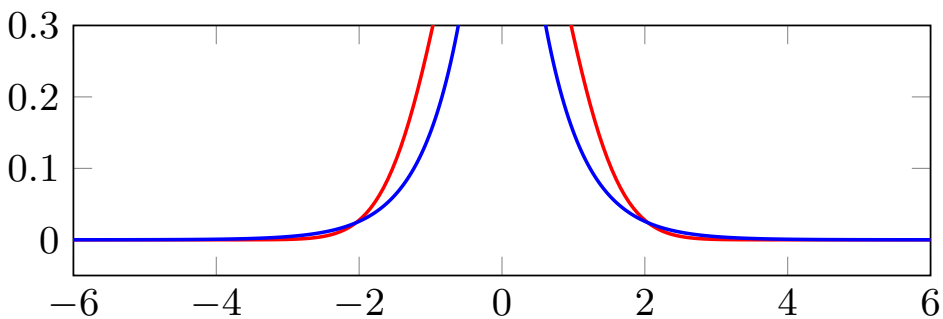
If the data has outliers, we can model the noise using a distribution that has heavier tails

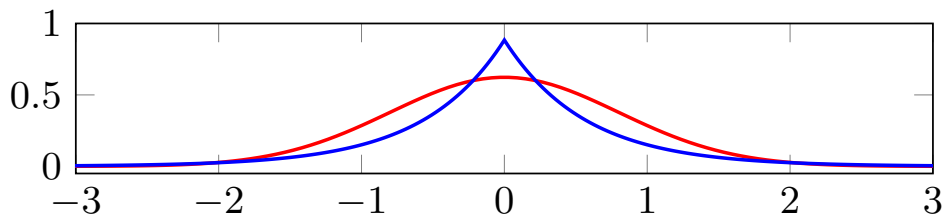
For the linear model $y = \mathbf{w} \cdot \mathbf{x} + \epsilon$, use

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

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

Given data $\langle(\mathbf{x}_i, y_i)\rangle_{i=1}^N$, let us express the likelihood of observing the data in terms model parameters $\mathbf{w}$ and b.

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

As in the case of the Gaussian noise model, we look at the negative log-likelihood

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

Thus, the maximum likelihood estimate in this case 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 fitting a linear model that is robust to outliers.

Outline

Probability Review

Linear Regression and Maximum Likelihood

Information, Entropy, KL Divergence

Lookahead: Binary Classification

Bernoulli random variable X takes value in $\{0, 1\}$. We parametrize using $\theta \in [0, 1]$.

$$p(1 \mid \theta) = \theta$$
$$p(0 \mid \theta) = 1 - \theta$$

More succinctly, we can write

$$p(x \mid \theta) = \theta^x (1 - \theta)^{1-x}$$

For binary classification problems, we will design models with parameter $\mathbf{w}$ that given input $\mathbf{x}$ produce a value $f(\mathbf{x}, \mathbf{w}) \in [0, 1]$. Then, we can model the (binary) class labels as:

$$y \sim \text{Bernoulli}(f(\mathbf{x}, \mathbf{w}))$$

Entropy

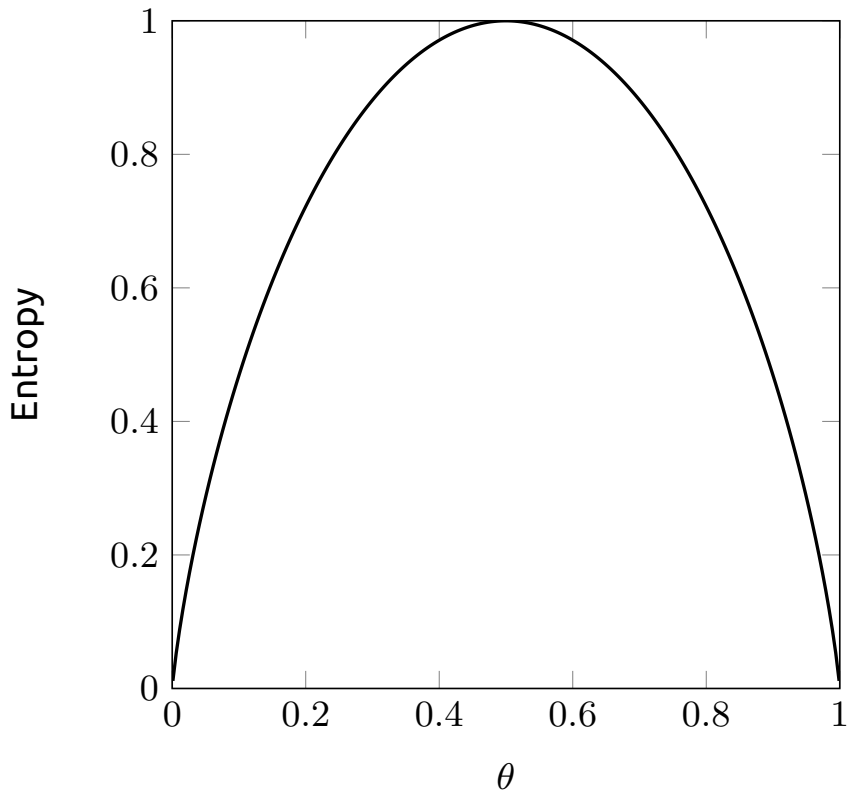
In information theory, entropy H is a measure of uncertainty associated with a random variable

$$H(X) = -\sum_{x} p(x) \log(p(x))$$

In the case of Bernoulli variables (with parameter θ) we get:

$$H(X) = -\theta \log_2(\theta) - (1-\theta) \log_2(1-\theta)$$

Entropy is a useful way to quantify information



Kullback-Leibler Divergence

KL-Divergence between distributions p and q is

$$\begin{aligned}\mathrm{KL}(p\|q) &= \sum_x p(x)\log\left(\frac{p(x)}{q(x)}\right) \\ &= \sum_x p(x)\log(p(x)) - \sum_x p(x)\log(q(x)) = -H(p) + H(p,q)\end{aligned}$$

$H(p,q) = -\sum_x p(x)\log(q(x))$ is called the cross-entropy

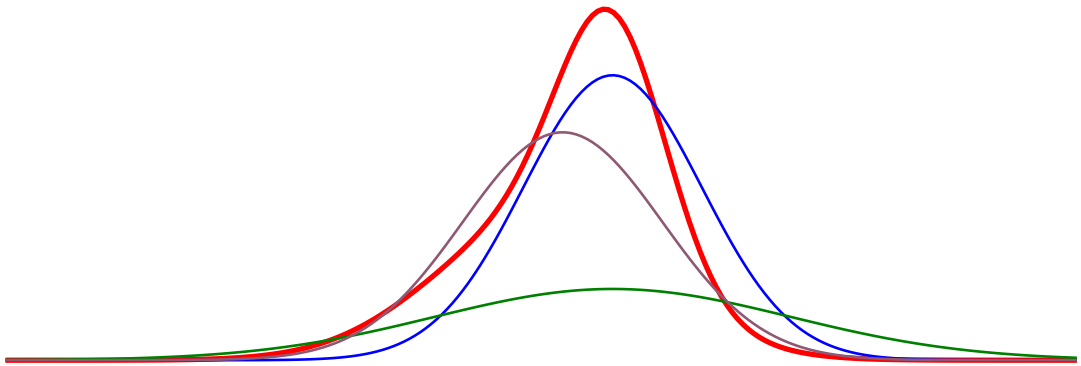
Properties of KL Divergence

- $\mathrm{KL}(p\|q) \geq 0$
- $\mathrm{KL}(p\|q) = 0$ if and only if $p = q$

Cross entropy accounts of the excess bits required to encode an observation from p if the encoding scheme was based on q

Maximum Likelihood and KL-Divergence

Suppose we get data $x_1, \ldots x_N$ from some unknown distribution p



Suppose we want to find the best parameters of a Guassian that explains the data, *i.e.*, we wish to estimate μ and σ.

We will find parameters that maximise the likelihood.

Maximum Likelihood and KL-Divergence

Suppose we get data $x_1, \ldots x_N$ from some unknown distribution p

Attempt to find parameters θ for distribution q from a family of distributions that best explains observed data

$$
\begin{aligned}
\widehat{\theta} &= \operatorname*{argmax}_{\theta} \prod_{i=1}^{N} q(x_i \mid \theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{i=1}^{N} \log(q(x_i \mid \theta)) \\
&= \operatorname*{argmax}_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log(q(x_i \mid \theta)) - \frac{1}{N} \sum_{i=1}^{N} \log(p(x_i)) \\
&= \operatorname*{argmin}_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{p(x_i)}{q(x_i \mid \theta)} \right) \\
&\to \operatorname*{argmin}_{\theta} \int \log \left(\frac{p(x)}{q(x|\theta)} \right) p(x) dx = \mathrm{KL}(p \| q_\theta)
\end{aligned}
$$

Next Time

- Beyond Linearity: Basis Expansion, Kernels
- Regularization: Ridge Regression, LASSO
- Overfitting, Model Complexity, Cross Validation