# Introduction

The goal of a Machine Learning (ML) algorithm is to find patterns in data, and make predictions. Usually we have a <u>training set</u> $\{x_1, \dots x_N\}$ and a model $y(x, w)$ where $w$ are the parameters. We fit $w$ against the training set obtaining $w^*$. Thus, given a new data point $x$, we can predict the value of a <u>target</u> variable $t$: $\hat{t} = y(x, w^*)$. The ability of our trained model to predict unseen data is called <u>generalization</u>. To obtain $w^*$ we minimize a <u>loss</u> or <u>error</u> <u>function</u> $E(w)$, i.e. $w^* = \arg\min_{w} E(w)$.

| $\{x_i\}$, $y(x, w)$ | $\longrightarrow$ | $y(x, w^*)$, $w^* = \arg\min E(w)$ | $\longrightarrow$ | $\hat{t} = y(x, w^*)$ |
|---|---|---|---|---|
| data + model | | learning | | prediction |

If each point in the training data is of the form $(x, t)$ we have <u>supervised</u> <u>learning</u>. If its only $x$, without target, its <u>unsupervised</u> <u>learning</u>. If $t$ is continuous, we have <u>regression</u>. If $t$ assumes a discrete or categorical value, we have <u>classification</u>. Its common to have a separate data set called <u>test</u> <u>set</u> (with targets) so we can evaluate the error of our trained model on unseen data to access generalization.

# Polynomial Fitting

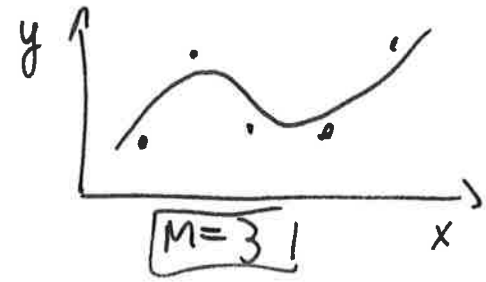
$X = (x_1, x_2, \dots, x_N)^T$
$T = (t_1, t_2, \dots, t_N)^T$

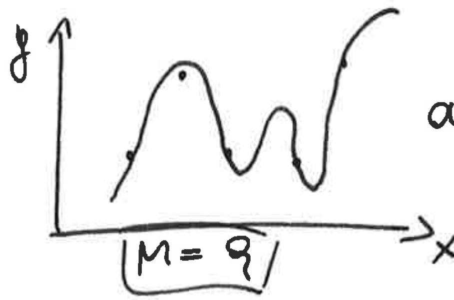
$$y(x, w) = w_0 + w_1 x + \dots + w_M x^M$$

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} \left( y(x_n, w) - t_n \right)^2$$

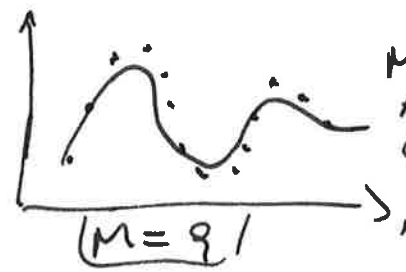
$$w^* = \arg\min_{w} E(w)$$

If $N > M$ we can fit the parameters.
$M$ is the complexity of our model.




overfit




more data, ALLOWS more COMPLEXITY.

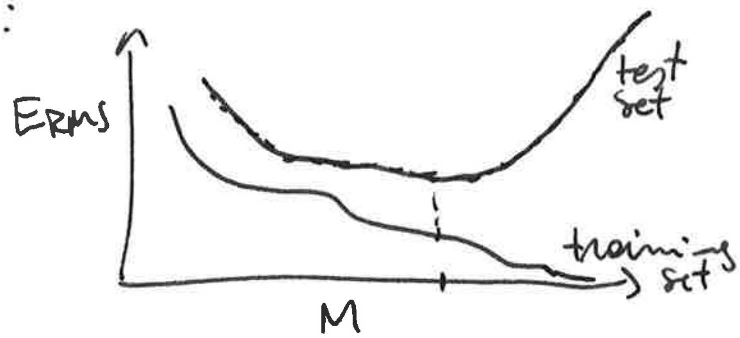
If the model is too complex, and we don't have enough data, the trained model will basically pass through each point. We are fitting noise in the data. Poor generalization.

However, if we have more data, then we can afford to a more complex model without overfitting.

we can have a separate test set $\{\tilde{x}_i\}, \{\tilde{t}_i\}$ and compute $\tilde{E} = \frac{1}{2} \sum_{i=1}^{L} \left( y(\tilde{x}_i, w^*) - \tilde{t}_i \right)^2$ as the test error. To make things of the same dimension as $t$, and independent of the data size, we use

$$E_{RMS} = \sqrt{\frac{2E(w^*)}{N}}$$

Tipically:

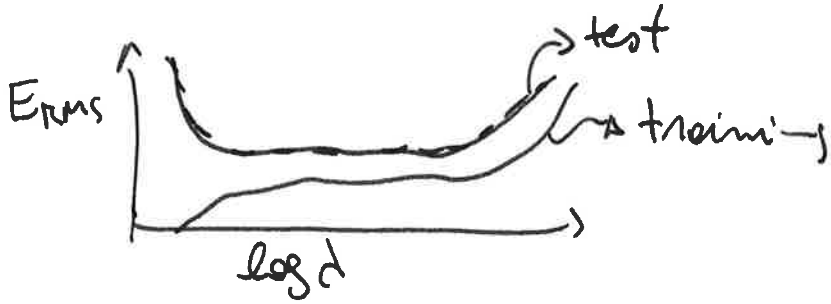



we can choose the right complexity.

When M increases and we don't have too much data, the value of $w_m$ is large. We can penalize the error function to control over fitting

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} \left( y(x_n, w) - t_n \right)^2 + \frac{\lambda}{2} \| w \|^2$$

(Ridge regression.) There is an optimal value of $\lambda$ which can be seen as




To make these ideas more systematic we need probability theory.

# Probability Theory

$$P(A) = \frac{N(A)}{N}$$

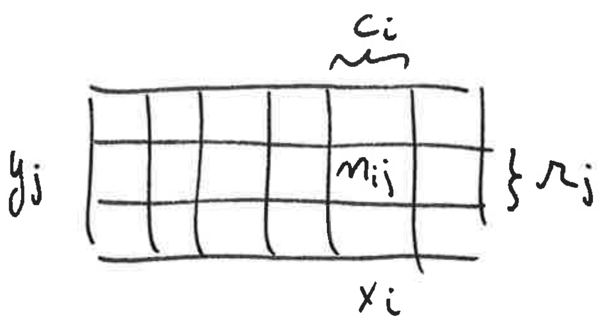
$P(A)$ ← event

$N(A)$ → Number occurrences of A

$N$ → total number of outcomes.

This is the frequentist view.

Two random variables: $X \in \{x_1, x_2, \dots, x_M\}$
$Y \in \{y_1, y_2, \dots, y_L\}$




$$P(x_i, y_j) = \frac{n_{ij}}{N}, \quad P(x_i) = \frac{c_i}{N}, \quad P(y_j) = \frac{r_j}{N}$$

$$c_i = \sum_j n_{ij}, \quad r_j = \sum_i n_{ij}$$

$$\boxed{\begin{aligned} P(x_i) &= \frac{c_i}{N} = \sum_j \frac{n_{ij}}{N} = \sum_j P(x_i, y_j) \\ P(y_j) &= \frac{r_j}{N} = \sum_i \frac{n_{ij}}{N} = \sum_i P(y_j, x_i) \end{aligned}}$$

Marginal probabilities

$$P(y_j \mid x_i) = \frac{n_{ij}}{c_i}$$

$$P(x_i \mid y_j) = \frac{n_{ij}}{r_j}$$

$$\cancel{P(y_j \mid x_i) = \frac{n_{ij}}{c_i} \cdot N}$$

→

$$\boxed{P(x_i, y_j) = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} = P(y_j \mid x_i)\, P(x_i)}$$

$$\boxed{P(x_i, y_j) = \frac{n_{ij}}{r_j} \cdot \frac{r_j}{N} = P(x_i \mid y_j)\, P(y_j)}$$

So we have
$$\begin{cases} p(X) = \sum_Y P(X,Y) & \text{(sum)} \\ p(X,y) = P(y|X)\,p(X) & \text{(product)} \end{cases}$$

Moreover,
$$\boxed{\underbrace{P(X|Y)}_{\text{posterior}} = \frac{\underbrace{P(Y|X)}_{\text{likelihood}}\ \underbrace{P(X)}_{\text{prior}}}{\underbrace{P(Y)}_{\text{Normalization}}}}$$
Bayes Theorem.

$$P(y) = \sum_X P(Y,X) = \sum_X P(y|X)\,P(X)$$

For a continuous variable:
$p(x)\,dx$ is the prob of $x \in (x, x+dx)$.

Thus $p(X \in [a,b]) = \int_a^b p(x)\,dx$

Conditions: $\begin{cases} p(x) \geq 0 \\ \int_{-\infty}^{\infty} p(x)\,dx = 1. \end{cases}$ $p$ is the prob. density

$P(\tilde{x}) = \int_{-\infty}^{\tilde{x}} p(x)\,dx$ is the cumulative probability

Change of variables: $x = g(y)$

$$p_x(x)\,dx = p_y(y)\,dy \quad \therefore \quad \boxed{p_y(y)} = p_x(x)\left|\frac{dx}{dy}\right| = p_x(g(y))\left|\frac{dg(y)}{dy}\right|$$

In higher dimensions
$$p_y(y) = p_x(g(y))\,|J| \quad \text{(Jacobian)}$$

sum $p(x) = \int p(x,y)\,dy$

prod. $p(x,y) = p(x|y)\,p(y)$

Expectation: $\mathbb{E}f(x) = \sum_x p(x)\,f(x)$ discrete

$\mathbb{E}f(x) = \int p(x)\,f(x)\,dx$ continuous.

If we have a sample $\{x_i\}_{i=1}^N$ then in either case

$$\mathbb{E}f(x) \approx \sum_{n=1}^{N} f(x_n)$$

$\{x_i\}$ must be drawn from $p(x)$!

$$\mathbb{E}_x f(x,y) = \int f(x,y)\,p(x)\,dx$$

$$\mathbb{E}_x[f(x) | y] = \int f(x)\,p(x|y)\,dx$$

$$\mathrm{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}f(x))^2]$$

$$= \mathbb{E}[f^2(x)] - (\mathbb{E}f(x))^2$$

$$\mathrm{cov}[x,y] = \mathbb{E}[(x - \mathbb{E}x)(y - \mathbb{E}y)^T]$$

## Bayesian Approach

Before: frequentist or classical. Repeatable experiments.
Bayes: only 1 data, uncertainty in the parameters.

posterior $\propto$ likelihood $\times$ prior

central role

Consider a gaussian distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$\beta = \frac{1}{\sigma^2}$ is the precision.

$\mathcal{N} > 0$, $\int_{-\infty}^{\infty} dx\, \mathcal{N}(x|\mu,\sigma^2) = 1$.

$$\mathbb{E}(x) = \int_{-\infty}^{\infty} x\, \mathcal{N}(x|\mu,\sigma^2)\, dx = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx$$

$$= \int_{-\infty}^{\infty} (\sigma y + \mu) \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}y^2} \sigma\, dy$$

$$= \sigma \underbrace{\int_{-\infty}^{\infty} \frac{y}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy}_{=0,\ \text{odd}} + \mu \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy}_{1}$$

$$\boxed{\mathbb{E}(x) = \mu}$$

$$\boxed{\mathbb{E}(x^2)} = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)^2} dx$$

$$= \int_{-\infty}^{\infty} (\sigma y + \mu)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

$$= \int_{-\infty}^{\infty} (\sigma^2 y^2 + \underbrace{2\sigma y \mu}_{\text{odd}} + \mu^2) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

$$= \mu^2 + \sigma^2 \underbrace{\int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy} = \boxed{\mu^2 + \sigma^2}$$

$$\frac{2}{\sqrt{2\pi}} \int_0^{\infty} y^2 e^{-\frac{y^2}{2}} dy = \frac{2}{\sqrt{2\pi}} \left(-2\frac{\partial}{\partial\beta}\right) \int_0^{\infty} e^{-\frac{\beta y^2}{2}} dy \Big|_{\beta=1}$$

$$= \frac{1}{\sqrt{2\pi}} \left(-2\frac{\partial}{\partial\beta}\right) \left(\sqrt{2\pi\beta}\right) = 1$$

$$\boxed{\sigma^2 = \mathbb{E}x^2 - (\mathbb{E}x)^2}$$

The point where $p(x)$ is maximum is the mode.
It's easy to see that this point is $x=\mu$.
It's possible to generalize this to higher dimensions:

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

Let $\mathbf{x} = (x_1, \dots, x_N)^T$ be an iid data set. The

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2) \qquad \text{likelihood.}$$

$$\log p = \sum_{n=1}^{N} \log \mathcal{N}(x_n|\mu,\sigma^2)$$

$$= \sum_n -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\left(\frac{x_n-\mu}{\sigma}\right)^2$$

$$l = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2$$

$$\frac{\partial l}{\partial \mu} = 0 \Rightarrow \boxed{\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n}$$

$$\frac{\partial l}{\partial \sigma} = 0 \Rightarrow -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{n=1}^{N}(x_n-\mu)^2 = 0$$

$$\boxed{\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\hat{\mu})^2}$$

$$\boxed{\mathbb{E}(\hat{\mu}) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\mathbb{E}(x_n)}_{\int_{-\infty}^{\infty} x\,\mathcal{N}(x)\,dx} = \mu}$$

~~iid sample~~ ~~iid sample~~

$$\cancel{\mathbb{E}(\hat{\sigma}^2) = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}(x_n^2) - 2\mathbb{E}(x_n)\mathbb{E}(\hat{\mu}) + \mathbb{E}[\hat{\mu}^2]}$$

$$\cancel{\mathbb{E}(\hat{\mu})^2}$$

$$\cancel{= \frac{1}{N}\sum_{n=1}^{N}(\mu^2+\sigma^2-2\mu^2+\mu^2)}$$

$$\mathbb{E}[\hat\sigma^2] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[x_n^2 - 2x_n\hat\mu + \hat\mu^2]$$

$$= \mathbb{E}[x_n^2] - \frac{2}{N}\sum_{n=1}^{N}\mathbb{E}[x_n\hat\mu] + \mathbb{E}[\hat\mu^2]$$

$$\mathbb{E}[x_n^2] = \sigma^2 + \mu^2$$

$$\sum_{n=1}^{N}\mathbb{E}\Big[x_n \underbrace{\frac{1}{N}\sum_{m=1}^{N} x_m}_{\hat\mu}\Big] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[x_n^2 + \Big(\sum_{\substack{m=1\\ m\neq n}}^{N} x_m\Big)x_n\Big]$$

$$= \mathbb{E}[x_n^2] + \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_m](N-1)\mathbb{E}(x_m)$$

$$= \mathbb{E}[x_n^2] + (N-1)(\mathbb{E}[x_m])^2$$

$$= \sigma^2 + \mu^2 + (N-1)\mu^2$$

$$= \sigma^2 + N\mu^2$$

$$\mathbb{E}[\hat\mu^2] = \frac{1}{N^2}\mathbb{E}\Big[\sum_{n=1}^{N}\sum_{m=1}^{N} x_n x_m\Big] = \frac{1}{N^2}\mathbb{E}\Big[\sum_{n=1}^{N} x_n^2 + \sum_{\substack{n=1\\ n\neq m}}^{N}\sum_{m=1}^{N} x_n x_m\Big]$$

$$= \frac{1}{N}\mathbb{E}[x_n^2] + \frac{1}{N^2}N(N-1)(\mathbb{E}(x_n))^2$$

$$= \frac{1}{N}(\sigma^2 + \mu^2) + \frac{1}{N}(N-1)\mu^2$$

$$= \frac{1}{N}\sigma^2 + \mu^2$$

Thus $\mathbb{E}[\hat\sigma^2] = \sigma^2 + \mu^2 - \frac{2}{N}(\sigma^2 + N\mu^2) + \frac{1}{N}\sigma^2 + \mu^2$

$$= \sigma^2\Big(1 - \frac{2}{N} + \frac{1}{N}\Big) + \mu^2(1 - 2 + 1)$$

$$\boxed{\mathbb{E}[\hat\sigma^2] = \frac{N-1}{N}\sigma^2}$$

So the Maximum Likelihood principle underestimate the variance (bias). For large $N$ this is Not a problem. An unbiased estimator is

$$\tilde{\sigma}^2 = \frac{N}{N-1}\hat{\sigma}^2 \Rightarrow \mathbb{E}[\tilde{\sigma}^2] = \sigma^2.$$

Thus $\tilde{\sigma}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$

## Curve Fitting

$t = y(x)$. on basis of training set

$$\mathbf{x} = (x_1, \dots, x_N)^T$$
$$\mathbf{t} = (t_1, \dots, t_N)^T$$

- We express the uncertainty over the target using a probability distribution.

$$p(t|x, w, \beta) = \mathcal{N}(t | y(x,w), \beta^{-1})$$

Likelihood function

$$p(\underset{\uparrow}{\mathbf{t}}|\underset{\uparrow}{\mathbf{x}}, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n | y(x_n, w), \beta^{-1})$$

training data

$$\log p = \sum_{n=1}^{N} \frac{1}{2}\log\beta - \frac{1}{2}\log 2\pi - \frac{1}{2}\beta(t_n - y(x_n,w))^2$$
$$= -\frac{\beta}{2}\sum_{n=1}^{N}(y(x_n,w) - t_n)^2 + \frac{N}{2}\log\beta - \frac{N}{2}\log 2\pi$$

Maximizing over $w$:

$$w^* = \arg\min_{w} \frac{1}{2}\sum_{n=1}^{N}(y(x_n,w) - t_n)^2$$ sum of squares error function

Maximizing with respect to $\beta$:

$$-\frac{1}{2}\sum_{n=1}^{N}(y(x_n,w^*)-t_n)^2+\frac{N}{2}\frac{1}{\beta}=0$$

$$\frac{1}{\beta^*}=\frac{1}{N}\sum_{n=1}^{N}(y(x_n,w^*)-t_n)^2$$

Now we can make predictions:

$$p(t|x,w^*,\beta^*)=\mathcal{N}(t|\,y(x_n,w^*),\beta^{*-1})$$

This is our model.

We can give one step further and introduce a prior distribution for $w$:

$$p(w|\alpha)=\mathcal{N}(w|0,\alpha^{-1}I)$$

$$=\left(\frac{\alpha}{2\pi}\right)^{\frac{M+1}{2}}e^{-\frac{\alpha}{2}w^Tw}$$

Remember that $y=w_0+w_1x+\dots+w_Mx^M$.

$\alpha$ is called an hyperparameters. Thus

$$p(w|x,t,\alpha,\beta)\propto p(t|x,w,\alpha,\beta)\,p(w|\alpha)$$

Now we can maximize the posterior (MAP).

$$\log p(w|x,t,\alpha,\beta)\propto\log p(t|x,w,\alpha,\beta)+\log p(w|\alpha)$$

$$\propto-\frac{\beta}{2}\sum_n(y(x_n,w)-t_n)^2-\frac{\alpha}{2}w^Tw$$ (keeping only $w$ terms)

$$w^*=\underset{w}{\operatorname{argmin}}\ \frac{1}{2}\sum_{n=1}^{N}(y(x_n,w)-t_n)^2+\frac{1}{2}\lambda\,w^Tw$$

where $\lambda=\frac{\alpha}{\beta}$. Regularized sum of squares.

# Bayesian curve fitting

We want $p(t \mid x, X, T, \alpha, \beta) = p(t \mid x, X, T)$

Assumed known ($\alpha, \beta$) — data ($X, T$).

Before we just did point estimation on $w$. For a full Bayesian treatment we must integrate over all $w$:

$$p(t \mid x, X, T) = \int p(t \mid x, w)\, p(w \mid X, T)\, dw \qquad (*)$$

we are ommiting dependencies on $\alpha, \beta$. Above

$$p(t \mid x, w) = \mathcal{N}(t \mid y(x, w), \beta^{-1})$$

$$p(w \mid X, T) = \frac{p(t \mid X, w)\, p(w \mid \alpha)}{C} \quad \text{normalization}$$

likelihood function $\prod_{n=1}^{N} \mathcal{N}(t_n \mid y(x_n, w), \beta^{-1})$

Since both distributions in $(*)$ are Gaussians, the integral will be a Gaussian, where the mean and variance will depend only on $\alpha, \beta$ and on the data $(X, T)$.

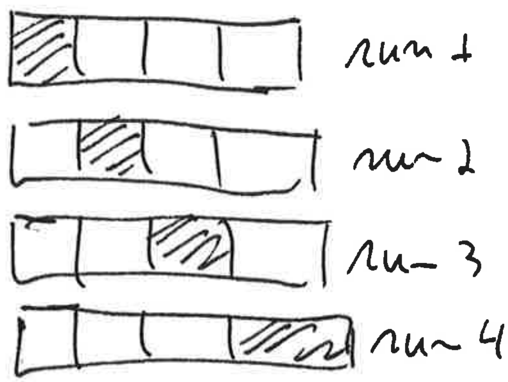
# Model Selection

We need to control the number of free parameters in our model, i.e. its complexity. We might also be interested in a range of different types of models in order to find the best one. This should be evaluated on untrained data.

If we have plenty of data we may split

| fit params | complexity | final evaluation |
| --- | --- | --- |
| training | validation | test |

If data is not plentiful we can use cross-validation.
We use a proportion $\frac{S-1}{S}$ for training, and $\frac{1}{S}$ for validation, to assess performance. When $S = N$ we have leave-one-out technique.




Average the results.

Drawbacks: more training runs, which can be expensive. When we use separate data to assess performance, and we have multiple complexity parameters, to test all combinations may require an exponential number of runs (in the # of params).

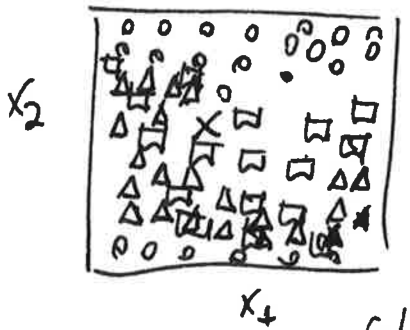
We need a better approach. Must rely only on the training data, and hyperparameters and model types must be compared on a single training run. We need a measure of performance that depends only on the training data, and does not over-fit.

Examples: AIC, BIC → variant of this. More later.

$$\max\left(\log p(D \mid w_{ML}) - M\right)$$

Curse of Dimensionality

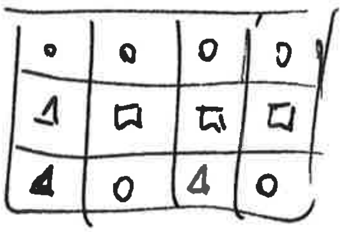
When data lies in a space of high dimension (many features) problems arise!




how to classify x?

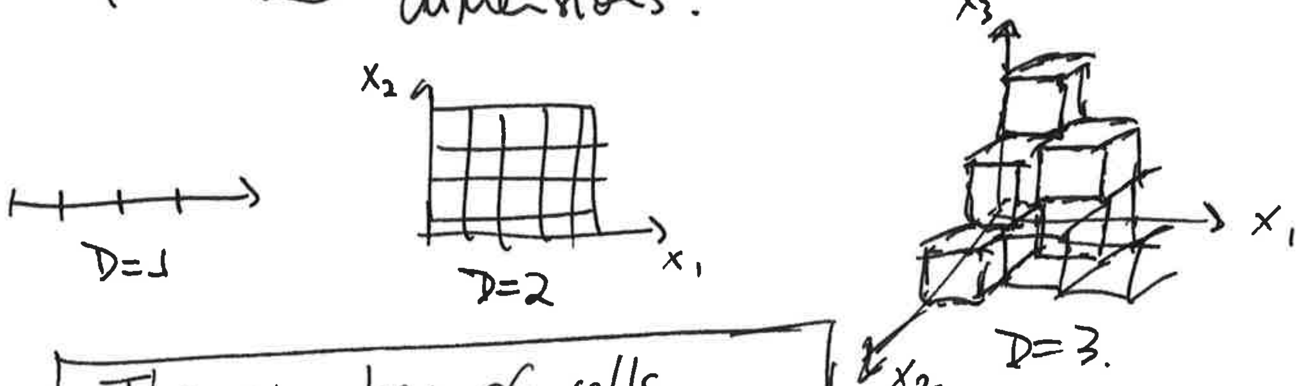
Nearest Neighbors is a reasonable assumption.

we can divide the space into cells an count the number of points in each cell.



so x would be a □.

Problem: what happens in higher dimensions?




$\boxed{\text{The number of cells grows exponentially with } D!}$

we thus need an exponentially larger amount of data to assure that no cell is empty.

Suppose we have a polynomial in D dimensions, i.e $x \in \mathbb{R}^D$. The

$$y(x,w) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k + \dots$$

The number of parameters increases drastically. For a polynomial of degree M, it is of $\mathcal{O}(D^M)$.

Another argument. Consider the volume of a sphere in D dimensions:

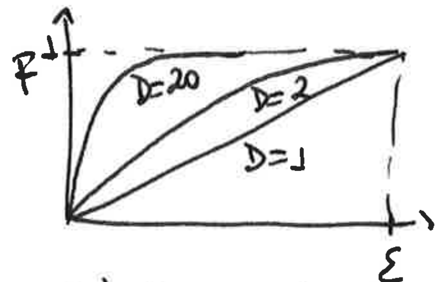
$$V_D(r) = k_D r^D$$

The fraction of the volume between $r=1-\varepsilon$ and $r=1$ is

$$F = \frac{V_D(1) - V_D(1-\varepsilon)}{V_D(1)} = 1 - \frac{k_D(1-\varepsilon)^D}{k_D} = 1 - (1-\varepsilon)^D$$

Making a plot:

when D increases, V increases so fast, that data become sparse!




In high dimensions, most of the volume of the sphere is concentrated in a thin shell close to its surface.

Also, in high dimensions, most of the probability mass of a gaussian is concentrated on a thin shell around a specific r.

In real data we can explore:

- usually data will fall in a lower effective dimensional subspace
- smoothness. $x \to x+\varepsilon \Rightarrow t \to t+\delta$. so we can explore some local interpolation to make predictions.

# Decision Theory

Suppose we have an input $x$ and a target $t$. Our goal is to predict $t$ for a new input $t$. The joint $p(x,t)$ provides a complete summary of the uncertainty between these variables. After inferring $t$ we can take a decision based on its value. Suppose $t$ is a two-class label, i.e. $t=0$ if $x \in C_1$ and $t=1$ if $x \in C_2$. The general problem consists in estimating $p(x, C_k)$. So given a new data $X$, we want $p(C_k, X)$, which through Bayes' theorem

$$P(C_k|X) = \frac{p(X|C_k)\,p(C_k)}{P(X)}$$

we want to maximize the posterior.

## Misclassification rate

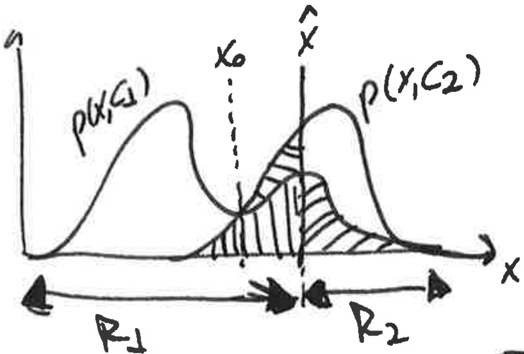
Rule that assigns $x$ to $C_k$. This rule divides the input space into regions $R_k$ called <u>decision regions</u>. All points in $R_k$ are assigned to $C_k$. The boundaries between $R_k$'s are called <u>decision boundaries</u>, or <u>decision surfaces</u>. Consider $\{C_1, C_2\}$ only.

$$\begin{aligned} p(\text{mistake}) &= p(X \in R_1, C_2) + p(X \in R_2, C_1) \\ &= \int_{R_1} p(x, C_2)\,dx + \int_{R_2} p(x, C_1)\,dx \end{aligned}$$

$$p(x, C_k) = p(C_k|X)\,p(x)$$

common factor.

To minimize the mistake, we want to maximize the correct posterior prediction $p(C_k|X)$.




The shaded regions correspond to mistakes.

$x \geq \hat{x} \Rightarrow x \in C_2$
$x < \hat{x} \Rightarrow x \in C_1$

- points from $C_2$ misclassified as $C_1$
- points from $C_2$ misclassified as $C_1$
- points from $C_1$ misclassified as $C_2$

No matter where $\hat{x}$ is and won't change. However we can change the area , so the best decision boundary is $\hat{x} = x_0$ where it vanishes.
This is the point where the curves $p(x, C_1)$ and $p(x, C_2)$ cross!

For $K$ classes, the probability of being correct is:

$$p(\text{correct}) = \sum_{k=1}^{K} p(x \in R_k, C_k)$$
$$= \sum_{k=1}^{K} \int_{R_k} \underbrace{p(x, C_k)}_{\hookrightarrow\, p(C_k|x)p(x)} dx$$

Again, this corresponds to $\max p(C_k|x)$.

## Expected Loss

Suppose $x$ is a patient and $C_1$ means he has cancer, and $C_2$ means he is healthy. If we classify $x \in C_1$ but actually $x \in C_2$, the implications are stress on the patient, and some collateral effect due to unnecessary drug administration. However, if we classify $x \in C_2$ but actually $x \in C_1$, the consequences are much more serious: premature death! So both types of errors are not equivalent.

we can formalize this through a loss function, or cost function.

true: $x \in C_k$ $\xrightarrow{\text{loss}}$ $L_{kj}$ element of a loss Matrix.
estimate: $x \in C_j$

$$L = \begin{array}{c|cc} & \text{cancer} & \text{healthy} \\ \hline \text{cancer} & 0 & 1000 \\ \text{healthy} & 1 & 0 \end{array}$$

← estimate.

↑ true

Attribute ≠ penalties.

The optimal solution is the one that minimizes the loss function. The uncertainty comes from $p(x, C_k)$. We minimize w.r.t the average:

$$\min_{\{R_j\}} \mathbb{E}[L] = \sum_k \sum_j \int_{R_j} L_{kj}\, p(x, C_k)\, dx$$

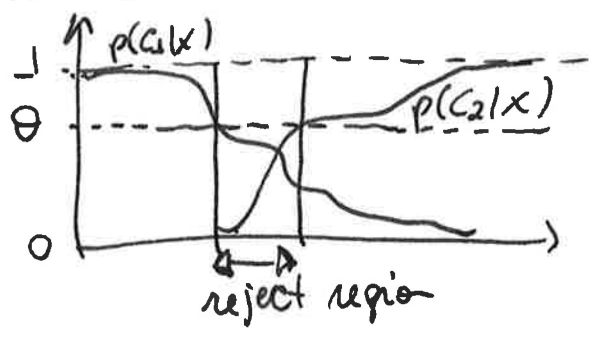
For each $x$ we should minimize $\sum_k L_{kj}\, p(x, C_k) \sim \boxed{\sum_k L_{kj}\, p(C_k|x)}$

Reject Option

We make mistakes when $p(x, C_i) \sim p(x, C_j)$, for $i \neq j$, or equivalently when $p(C_k|x) \ll 1$. We can introduce a threshold variable $\theta$ such that if

$$\max\{p(C_i|x), p(C_j|x)\} \leq \theta$$

we don't take any action, or we reject any prediction, which should then be more carefully analyzed by a better method.



If $\theta = 1$ we reject all. If $\theta < 1/K$ we don't reject any.

Inference and Decision

So far we have { 1. inference stage: $p(C_k|X)$ training data
2. decision stage: use these posteriors to optimize assignment.

An alternative would be to do everything together directly using the training data: $f: X \to C_k$. $f$ is a discriminant function (This approach is not usually recommended).

3 Approaches:

(a) Solve inference problem $p(X|C_k)$. Infer $p(C_k)$. Then use $p(C_k|X) = \dfrac{p(X|C_k)\,p(C_k)}{\sum_k p(X|C_k)\,p(C_k)}$.

This is equivalent to infer $p(X, C_k)$. After this we can use decision theory. Approaches that model distr. of inputs as well as distr. of outputs are known as generative models.
"This method is expensive and complex."

(b) Solve the inf. prob. $p(C_k|X)$. Then use decision theory. Approaches that model the posterior are known as discriminative models. "Less expensive, not so powerfull."

(c) find discriminative function $f(x)$ from the training data. Can be bad, especially if "we want to change something later."

# Loss Function For Regression

We want to choose a function $y(x)$ to predict $t$, based on a loss function $L(y(x),t)$, such as to minimize the expected loss

$$\mathbb{E}[L] = \iint dx\,dt\; p(x,t)\,L(y(x),t)$$

For fixed $(x,t)$ we vary $y \to y + \delta y$ thus

$$\delta L = \frac{\partial L}{\partial y}\,\delta y$$ to leading order. We thus find $y$ by solving

$$\frac{\delta \mathbb{E}[L]}{\delta y} = 0 = \int dt\; p(x,t)\,\frac{\partial L}{\partial y}$$

For squared loss $L = (y-t)^2$ we have

$$0 = \int dt\; p(x,t)\,2(y(x)-t) \;\therefore\; \int dt\,p(x,t)\,y(x) = \int dt\; t\,p(x,t)$$

$$y(x)\cancel{\int dt\; p(x,t)} = \int dt\; t\; p(t|x)\cancel{p(x)}$$

$$\boxed{y(x) = \int dt\; t\; p(t|x) = \mathbb{E}_t[t|x]}$$

The optimal solution is the conditional average.

3 Approaches (a) Determine $p(x,t) \to p(t|x) \to \mathbb{E}_t[t|x]$

(b) Infer $p(t|x) \to \mathbb{E}_t[t|x]$

(c) Find $y(x)$ from $\{x\}$.

"In order of complexity more $\to$ less".

A more general loss functional would be

$$\mathbb{E}[L_q] = \int dx \int dt \, | y(x) - t |^q \, p(x,t)$$

Minkowski loss.

## Information Theory

How much information we gain after observing the value of a given random variable $x$?
If this value is unlikely, low prob, we gain a lot of information. If the value has high prob. the low information. If the value is certain to occur, then no information at all. We look for a monotonic function $h(\cdot)$ of $p(x)$ that express this information content.

if $x$ and $y$ are unrelated, observing both should give $h(x,y) = h(x) + h(y)$. Two unrelated events satisfy $p(x,y) = p(x)p(y)$. Thus $h(x) \propto \log p(x)$, or since $p \le 1$:

$$\boxed{h(x) = -\log p(x)}$$

$$\left.\begin{array}{l} p \text{ small} \Rightarrow h \text{ large} \\ p \text{ large} \Rightarrow h \text{ small} \end{array}\right\} \text{ as desired!}$$

Now suppose we send a set of messages $\{h(x)\}$.
The average ammount of information transmited is

$$\boxed{H(x) = -\sum_x p(x) \log p(x)}$$

which is the entropy. $p \to 0 \Rightarrow p \log p \to 0$

Noiseless coding theorem (Shannon 1948): entropy is a lower bound on the number of bits needed to transmit the state of a random variable.

Stat. Mech. View: $N$ objects to be divided into $N$ bins, such that there are $n_j$ objects in the $j$th bin. The number of ways we can do this is:

$$W = \frac{N!}{n_1!\, n_2! \cdots n_k!}$$ (Number of microstates)

Then, $H = \frac{1}{N} \log W = \frac{1}{N} \log N! - \frac{\sum_i \log n_i!}{N}$

Consider $N \to \infty$, but $\frac{n_i}{N}$ fixed. Stirling

$\log N! \cong N \log N - N$ Thus,

$$H \approx \log N - 1 - \frac{1}{N} \sum_i \log n_i!$$

$$\approx \log N - \frac{1}{N} \sum_i (n_i \log n_i - n_i)$$

$$\approx \log N - \sum_i \frac{n_i}{N} \log n_i$$

$$= \frac{1}{N} \sum n_i \log N - \sum_i \frac{n_i}{N} \log n_i = - \sum_i \frac{n_i}{N} \log \frac{n_i}{N}$$

$$= - \sum_i p_i \log p_i$$

The bins can be interpreted as the states of a random variable $X$ where $P(X = x_i) = p_i$. The entropy of a r.v. $X$ is thus

$$H[p] = - \sum_i p(x_i) \log p(x_i)$$

minimum value $H = 0$ when $p_i = 1$ and all $p_j = 0$ for $j \neq i$.

The maximum entropy can be found by solving:

$$\tilde{H} = \max_{p_i}\underbrace{\left(-\sum_i p_i \log p_i + \lambda\left(\sum_i p_i - 1\right)\right)}_{L}$$

$\lambda$ ← lagrange multiplier.

$$\frac{\partial L}{\partial p_j} = -\log p_j - 1 + \lambda = 0 \therefore \lambda - 1 = \log p_j$$

$$e^{\lambda-1} = p_j$$

$$\sum_j e^{\lambda-1} = \sum_j p_j$$

$$M e^{\lambda-1} = 1 \therefore e^{\lambda-1} = \frac{1}{M}$$

where $M$ is the # states. Thus $\boxed{p_j = \frac{1}{M}}$

$$\tilde{H} = -\sum_i \frac{1}{M}\log\frac{1}{M} = -\log\frac{1}{M} = \boxed{\log M}$$

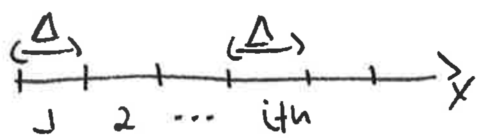
We can check that the second derivatives are negative:

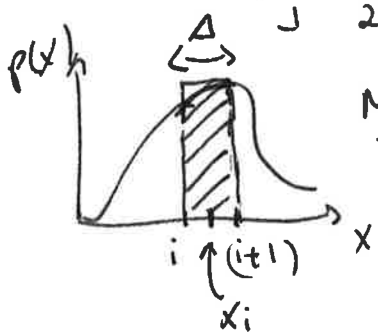
$$\frac{\partial \tilde{H}}{\partial p_j \partial p_i} = -\frac{1}{p_i}\delta_{ij}$$

so its a maximum.

uniform distr. gives the maximum entropy

Let us consider the continuous case. Divide $x$ into bins of size $\Delta$:




Mean value Theorem $\Rightarrow$ $\int_{i\Delta}^{(i+1)\Delta} p(x)dx = p(x_i)\Delta$

for some $x_i$ in the ith bin.

we can quantize $x$ by assigning the value $x_i$ whenever $x$ falls in the ith bin. The prob. of observing $x_i$ is then $p(x_i)\Delta$. Thus

$$H_\Delta = -\sum_i p(x_i)\Delta \log(p(x_i)\Delta) = -\sum_i p(x_i)\Delta \log p(x_i) - \underbrace{\left(\sum_i p(x_i)\Delta\right)}_{=1}\log\Delta$$

$$H_\Delta = -\sum_i p(x_i)\Delta \log p(x_i) - \log\Delta$$

& diverges when $\Delta \to 0$. Associated to the fact that we need infinite ammount of information to specify the state of a continuous variable. So we drop this term.

continuous limit:

$$\sum_i \Delta \to \int dx$$
$$x_i \to x$$

$$\boxed{H = -\int p(x)\log p(x)\,dx}$$ differential entropy.

Maximum entropy for a continuous distr.?

Constraints:
- $\int_{-\infty}^{\infty} p(x)dx = 1$ (normalization)
- $\int_{-\infty}^{\infty} x\,p(x)dx = \mu$ (first moment)
- $\int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \sigma^2$ (second moment)

Lagrangian:

$$L = -\int p\log p\,dx + \lambda_1\left(\int p\,dx - 1\right) + \lambda_2\left(\int x\,p\,dx - \mu\right) + \lambda_3\left(\int (x-\mu)^2 p\,dx - \sigma^2\right)$$

Making $p \to p + \delta p$ and keeping leading order terms yields

$$\frac{\delta L}{\delta p} = -\log p - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 = 0$$

$$p(x) = e^{-1+\lambda_1+\lambda_2 x+\lambda_3(x-\mu)^2}$$

Now we determine the Lagrange multipliers by putting this back into the constraints:

$$\int p\,dx = 1 = e^{-1+\lambda_1+\mu\lambda_2-\frac{\lambda_2^2}{4\lambda_3}} \sqrt{\frac{\pi}{(-\lambda_3)}}$$

$$\int x\,p\,dx = \mu = e^{-1+\lambda_1+\mu\lambda_2-\frac{\lambda_2^2}{4\lambda_3}} \sqrt{\frac{\pi}{-\lambda_3}} \frac{1}{(-2\lambda_3)} (\lambda_2 - 2\mu\lambda_3)$$

$$\int (x-\mu)^2 p\,dx = \sigma^2 = e^{-1+\lambda_1+\mu\lambda_2-\frac{\lambda_2^2}{4\lambda_3}} \sqrt{\frac{\pi}{-\lambda_3}} \frac{1}{(4\lambda_3^2)} (\lambda_2^2 - 2\lambda_3)$$

Thus:

$$\mu = \frac{1}{-2\lambda_3}(\lambda_2 - 2\mu\lambda_3)$$

$$\sigma^2 = \frac{1}{4\lambda_3^2}(\lambda_2^2 - 2\lambda_3)$$

$$\mu = -\frac{\lambda_2}{2\lambda_3} + \cancel{\frac{2\mu\lambda_3}{2\lambda_3}} \therefore -\frac{\lambda_2}{2\lambda_3} = 0 \therefore \boxed{\lambda_2 = 0}$$

$$\sigma^2 = \frac{1}{4\lambda_3^2}(-2\lambda_3) = -\frac{1}{2\lambda_3} \therefore \boxed{\lambda_3 = -\frac{1}{2\sigma^2}}$$

$$1 = e^{-1+\lambda_1} \sqrt{\frac{\pi}{(1/2\sigma^2)}} = e^{-1+\lambda_1}\sqrt{\pi\, 2\sigma^2}$$

$$\boxed{e^{-1+\lambda_1} = \frac{1}{\sqrt{2\pi\sigma^2}}}$$

Therefore, the distr. that maximizes the entropy is the normal distr,

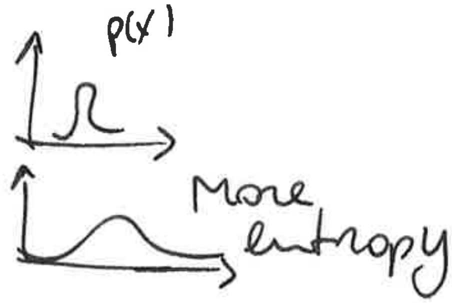
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Replacing this into $H[x]$:

$$H[x] = -\int dx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left( -\frac{1}{2}\log 2\pi\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2} \right)$$

$$= \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{2\sigma^2} \underbrace{\int dx\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} (x-\mu)^2}_{\sqrt{2\pi}\cdot\sigma^3}$$

$$\boxed{H[x] = \frac{1}{2}\left(1 + \log 2\pi\sigma^2\right)}$$




- $H$ increases if $\sigma$ increases!
- $H$ can be $< 0$ if $2\pi\sigma^2 < \frac{1}{e}$!

Suppose now we have the joint distr. $p(x,y)$. If $x$ is known, the amount of information needed to specify $y$ is $-\log p(y|x)$. Thus

$$H[y|x] = -\iint p(x,y)\log p(y|x)\,dx\,dy$$

Conditional entropy. Replacing $p(x,y) = p(y|x)p(x)$ we have

$$H[y|x] = -\iint p(x,y)\log p(x,y) + \iint p(x,y)\log p(x)$$

$$= H[x,y] - H[x]$$

$$\boxed{H[x,y] = H[y|x] + H[x]}$$

# Relative Entropy and Mutual Information

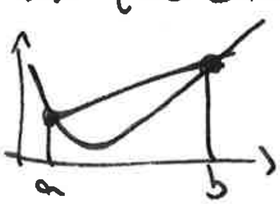
Unknown distribution $p(x)$.
Approximate distribution $q(x)$, intended to model $p(x)$. If we use $q(x)$ to transmit the "messages" the additional amount of information, ON Average, is

$$KL(p\|q) = -\int p(x)\log q(x)\,dx - \left(-\int p(x)\log p(x)\,dx\right)$$
$$= -\int p(x)\log\left(\frac{q(x)}{p(x)}\right)dx$$

Relative entropy or Kullback-Leibler divergence.
Not symmetric!

$\boxed{\text{We now show that } KL(p\|q) \geqslant 0, \text{ and the equality iff } p=q.}$

Convex functions: $f(\lambda a+(1-\lambda)b) \leqslant \lambda f(a)+(1-\lambda)f(b)$




Strictly convex if equality is satisfied only with $\lambda=0$ and $\lambda=1$.

This can be generalized to:

$$f\left(\sum_i \lambda_i x_i\right) \leqslant \sum_{i=1} \lambda_i f(x_i)$$

for $\lambda_i \geqslant 0$ and $\sum_i \lambda_i = 1$. This implies

$$f(\mathbb{E}[x]) \leqslant \mathbb{E}[f(x)]$$ Jensen's inequality

For the continuous case:

$$f\left(\int p(x)x\,dx\right) \leqslant \int f(x)p(x)\,dx$$

Thus

$$KL(p\|q) = -\int p \log \frac{q}{p}\, dx \geq -\log \underbrace{\int p \frac{q}{p}\, dx}_{=1} = 0$$

$\therefore KL(p\|q) \geq 0.$

$-\log$ is a convex function, actually strictly convex, thus equality implies $p = q$.

KL is a measure of dissimilarity between $p$ and $q$.

Suppose we model the unknown $p(x)$ by $q(x|\theta)$ and we wish to determine $\theta$.
We can try to minimize KL, however $p(x)$ is unknown. Now suppose we have an iid sample drawn from $p(x)$, $\{x_i\}_{i=1}^N$. Then

$$KL(p\|q) \approx -\sum_{i=1}^{N} \left(\log q(x_i|\theta) - \log p(x_i)\right)$$

does not depend on $\theta$

$$\begin{aligned}\theta^* &= \arg\min_\theta KL(p\|q(\theta)) \\ &= \arg\max_\theta \sum_{i=1}^{N} \log q(x_i|\theta) \\ &= \arg\max_\theta \prod_{i=1}^{N} q(x_i|\theta)\end{aligned}$$

Minimizing KL is equivalent to maximum likelihood

Consider $p(x,y)$. If $x \perp y$ the $p(x,y) = p(x)p(y)$. 29
We can measure in-dependence by

$$I[x,y] = KL(p(x,y) \| p(x)p(y))$$
$$= -\iint dx\, dy\; p(x,y) \log \frac{p(x)p(y)}{p(x,y)}$$

which is the mutual information. $I[x,y] \geq 0$.
with equality iff $x \perp y$.

We can write

$$p(x,y) \log \frac{p(x)p(y)}{p(x,y)} = p(x,y) \log p(x) + p(x,y) \log \frac{p(y)}{p(x,y)}$$
$$= p(x,y) \log p(x) - p(x,y) \log p(x|y)$$

or

$$= p(x,y) \log p(y) - p(x,y) \log p(y|x)$$

Thus

$$I[x,y] = H[x] - H[x|y]$$
$$= H[y] - H[y|x]$$

I is the reduction in the uncertainty about x after observation of y (and vice-versa).

$$\left.\begin{array}{l} p(x) \to \text{prior} \\ p(x|y) \to \text{posterior.} \end{array}\right\} \text{Bayesian interp.}$$