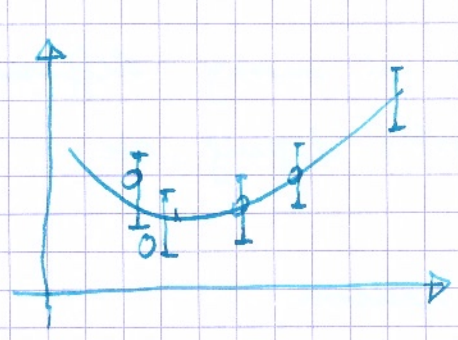
Bayesian Reasoning:

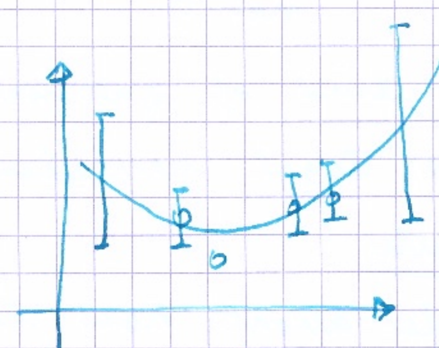
Consider the following model: $\hat{Y}_i = \hat{\theta}_0 + x_i \hat{\theta}_1 + x_i^2 \hat{\theta}_2$

Recall the probabilistic view of linear regression

$$y_i = \theta_0 + x_i \theta_1 + x_i^2 \theta_2 + \mathcal{N}(0, \sigma^2)$$

We are able to estimate $\theta_0, \theta_1, \theta_2$ using MLE estimation. We assume that every point has the same variance. And we are able to compute that variance using MLE.





This is what we want. Model not only returns an answer, it also returns a confidence in that answer.

- The ability to model uncertainity is very important. It also enables us to do the exploration vs. exploitation trade off.

  Example: drilling for minerals.

$\boxed{\text{Bayesian learning allows us to do just that.}}$

Problem: A doctor has a good news and a bad news for a patient.

Bad News: The patient has just tested positive for a serious disease. The test is 99% accurate.

Good News: This is a very rare disease. Only 1 in 10000 people have it.

Question: Should the patient be very worried or not?

* This kind of question arise in many different areas: legal proceedings, sports doping, etc.

Bayes Rule

- A rule about how humans reason
- TED Talk by Alex Gopnik (Psychologist at Berkley)
- This rule allows us to invert probabilities.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- Very useful for the inverse problems: computer vision
- key to perception.

$$\underbrace{P(AB)}_{\text{joint}} = P(B|A)\,\underbrace{P(A)}_{\text{marginal}} = P(A|B)\,\underbrace{P(B)}_{\text{marginal}}$$

Conditional

$P(B|A)$: prob. of B given A.

- $\int P(AB)\, dA\, dB = 1$

- $\int \underbrace{P(A|B)}\, \cancel{P(B)}\, dA = 1$

This is a distribution over A. B is given.

- $\int P(A)\, dA = 1$

- $\int P(AB)\, dB = P(A)$ ← integrating out B. this is also referred to marginalisation in the world of probability.

Learning and Bayesian Inference.

h: hypothesis

d: data

$$p(h|d) = \frac{p(d|h)\, p(h)}{\sum_{h' \in H} p(d|h')\, p(h')}$$

H: set of all possible hypothesis.

$p(h)$ refers to prior belief.

$p(d|h)$ refers to likelihood.

$p(h|d)$ refers to posterior.

$$p(h|d) = \frac{p(d|h)\, p(h)}{\sum_{h' \in H} p(d|h')\, p(h')} = \frac{p(d|h)\, p(h)}{p(d)}$$

This integral is very hard to do in practice. We can do it for ~~g~~ Gaussians though.

Lets use Bayesian rule to see if the patient should be worried.

1. Test is 99% accurate: $P(T=1 \mid D=1) = 0.99$
   and $P(T=0 \mid D=0) = 0.99$
2. Disease effects 1 in 10000 : $P(D=1) = 0.0001$

$$P(D=1 \mid T=1) = \frac{P(T=1 \mid D=1)\, P(D=1)}{P(T=1 \mid D=0)\, P(D=0) + P(T=1 \mid D=0)\, P(D=1)}$$

$$= \frac{(0.99)(0.0001)}{(1-0.99)(1-0.0001) + (0.99)(0.0001)}$$

$$= 0.0098$$

This is the probability of the patient having the disease given that the test came positive.

<u>Example:</u>

$$P(\text{words} \mid \text{sounds}) \propto \underbrace{P(\text{sound} \mid \text{words})}_{\text{likelihood of data.}}\ \underbrace{P(\text{words})}_{\text{prior language model.}}$$

unigram
bigram
etc.

Bayesian learning for Model Parameters

1. Given n data $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$, write down the likelihood of data $P(D \mid \theta)$
2. Specify prior $P(\theta)$.
3. Compute posterior $P(\theta \mid D) = \dfrac{P(D \mid \theta)\, P(\theta)}{P(D)}$

Prior $p(\theta)$ encodes our belief about the parameters.

We are assuming that θ is a random variable. Note that for MLE, θ is <u>not</u> a random variable.

Prior can also be viewed as my initial belief.

$$P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)} \propto P(D|\theta)\,P(\theta)$$

This can be seen ~~is~~ as simply normalizing $P(D|\theta)P(\theta)$.

* Frequentist vs. Bayesian : Bayesian claim that it is possible to assign probabilities without ever seeing the frequencies of events in question.

<u>Bayesian Linear Regression</u>

1. The likelihood is Gaussian $\mathcal{N}(y|X\theta, \sigma^2 I_n)$

   The conjugate prior is ~~as~~ also a Gaussian $P(\theta) = \mathcal{N}(\theta|\theta_0, V_0)$. θ_0 is the mean and V_0 is the variance.

$$p(\theta|X, y, \sigma^2) \propto \mathcal{N}(\theta|\theta_0, V_0)\,\mathcal{N}(y|X\theta, \sigma^2 I_n) = \mathcal{N}(\theta|\theta_n, V_n)$$

$$\theta_n = V_n V_0^{-1}\theta_0 - \frac{1}{\sigma^2} V_n X^T y$$

$$V_n^{-1} = V_0^{-1} - \frac{1}{\sigma^2} X^T X$$

This sort of calculations are also called "completing the squares" or conjugate analysis. Both prior and the posterior has the same shape.

* You need a course on "Bayesian Analysis"!

$$P(\theta \mid Y, X, \sigma^2) \propto P(Y \mid X, \theta, \sigma^2)\, P(\theta)$$

$$\propto e^{-\frac{1}{2\sigma^2}(Y-\theta X)^T(Y-X\theta)}\; e^{-\frac{1}{2}(\theta-\theta_0)^T V_0^{-1}(\theta-\theta_0)}$$

$$\cancel{= e^{-\frac{1}{2}\{(Y-X\theta)^T \sigma (Y-X\theta) + (\theta-\theta_0)^T V_0^{-1}}}$$

$$= e^{-\frac{1}{2}\{(Y-X\theta)^T \sigma^{-2}(Y-X\theta) + (\theta-\theta_0)^T V_0^{-1}(\theta-\theta_0)\}}$$

$$= e^{-\frac{1}{2}\{\underline{\sigma^{-2}Y^TY} - 2\sigma^{-2}Y^TX\theta + \underline{\sigma^{-2}\theta^TX^TX\theta} + \theta^TV_0^{-1}\theta \underline{- 2\theta_0^TV_0^{-1}\theta} + \theta_0^TV_0^{-1}\theta_0\}}$$

$$\cancel{= e^{-\frac{1}{2}\{\text{const} + (\sigma^{-2}\theta^TX^TX\theta - 2\theta_0^TV_0^{-1}\theta) + (}}$$

$$= e^{-\frac{1}{2}}\{(\sigma^{-2}Y^TY + \theta_0^TV_0^{-1}\theta_0)$$

$$(\sigma^{-2}\theta^TX^TX\theta + \theta^TV_0^{-1}\theta)$$

$$(-2\sigma^{-2}Y^TX\theta + \theta_0^TV_0^{-1}\theta + \cancel{\theta_0^TV_0^{-1}\theta_0})$$

$$= e^{-\frac{1}{2}}\{$$

$$\theta^T \underbrace{\left(X^T(\sigma^2 I)^{-1}X + V_0^{-1}\right)}_{V_n^{-1}}\theta$$

$$-2\left(\frac{Y^TX}{\sigma^2} + \theta_0^TV_0^{-1}\right)\theta$$

$$= \theta^T V_n^{-1}\theta - 2\theta_n^T V_n^{-1}\theta + 2\theta_n^T V_n^{-1}\theta$$

$$2\left[\theta_n^T V_n^{-1} - \frac{Y^TX}{\sigma^2} - \theta_0^T V_0^{-1}\right]\theta$$

Bayesian Linear Regression:

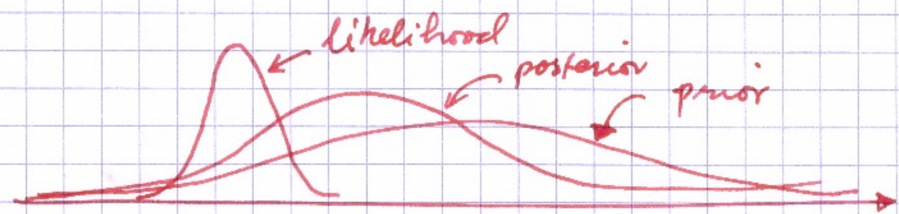
The likelihood is Gaussian: $\mathcal{N}(y \mid X\theta, \sigma^2 I_n)$

The conjugate prior is also Gaussian: $p(\theta) = \mathcal{N}(\theta \mid \theta_0, V_0)$

mean: θ_0

variance: V_0

Later you'll notice that if we make $\theta_0 = 0$ and V_0 equal to a diagonal matrix, what you get is a ridge regression, i.e., Bayesian Linear Regression will subsume Ridge Regression.



For ~~[illegible]~~ Bayesian Linear Regression, we will not compute a single value of θ. Instead we will estimate a distribution over θ. Specifically we want to compute the posterior $p(\theta \mid X, y, \sigma^2)$

$$p(\theta \mid X, y, \sigma^2) \propto \mathcal{N}(\theta \mid \theta_0, V_0)\, \mathcal{N}(y \mid X\theta, \sigma^2 I) = \mathcal{N}(\theta \mid \theta_n, V_n)$$

θ_n : mean

V_n : variance that models the uncertainty.

Through conjugate analysis or completing squares exercise we can find out the value ~~of~~ for θ_n and V_n.

$$\left.\begin{aligned} \theta_n &= V_n V_0^{-1} \theta_0 - \frac{1}{\sigma^2} V_n X^T y \\ V_n^{-1} &= V_0^{-1} - \frac{1}{\sigma^2} X^T X \end{aligned}\right] \text{Posterior.}$$

Conjugate analysis: what conjugate analysis means that the prior and the posterior has the shape.

* You ~~can~~ need a course on Bayesian statistics.

We are exploiting conjugate analysis. That is to say we are picking a prior such that our posterior has the same shape as the prior. Ideally I would like the freedom to pick any prior; however, that makes <u>analysis</u> / <u>computation</u> for posterior very difficult.

$\boxed{\text{Conjugate analysis}}$ * assume σ^2 is known.

$$P(\theta \mid Y, X, \sigma^2) \propto P(Y \mid X, \theta, \sigma^2)\, P(\theta)$$

$$\propto e^{-\frac{1}{2\sigma^2}(Y - X\theta)^T (Y - X\theta)}\, e^{-\frac{1}{2}(\theta-\theta_0)^T V_0^{-1} (\theta - \theta_0)}$$

Let's combine these two terms and complete squares.

Proportionality allows us to get rid of constant.

$$= e^{-\frac{1}{2}\{(Y - X\theta)^T \sigma^{-2} (Y - X\theta) + (\theta - \theta_0)^T V_0^{-1} (\theta - \theta_0)\}}$$

$$= e^{-\frac{1}{2}\{\sigma^{-2} \underline{Y^T Y} - 2\sigma^{-2} Y^T X \theta + \underline{\sigma^{-2} \theta^T X^T X \theta} + \underline{\theta^T V_0^{-1} \theta} - 2\theta_0^T V_0^{-1} \theta + \underline{\theta_0^T V_0^{-1}} \theta_0\}}$$

$$\sigma^{-2} \theta^T X^T X \theta + \theta^T V_0^{-1} \theta$$

$$= \theta^T (X^T (\sigma^2 I)^{-1} X + V_0^{-1}) \theta$$

$$= \theta^T V_n^{-1} \theta$$

$$= e^{-\frac{1}{2}\{\underline{\text{const}} + \theta^T V_n^{-1} \theta - 2\left(\frac{Y^T X}{\sigma^2} + \theta_0^T V_0^{-1}\right)\theta\}}$$

$$= e^{-1/2\{\text{const} + \theta^T V_n^{-1} \theta - 2\theta_n^T V_n^{-1} \theta + 2\theta_n^T V_n^{-1} \theta - 2\left(\frac{Y^T X}{\sigma^2} + \theta_0^T V_0^{-1}\right)\theta\}}$$

$$= e^{-1/2\{\text{const}_2 + (\theta - \theta_n)^T V_n^{-1} (\theta - \theta_n) + 2\underline{\left[\theta_n^T V_n^{-1} - \frac{Y^T X}{\sigma^2} - \theta_0^T V_0^{-1}\right]}\theta\}}$$

Doesn't matter $\because$ I am working upto a proportionality

We want to get rid of this term.

If we choose $\theta_0 = 0$ and $V_0 = \tau_0^2 I_d$, which is a spherical Gaussian prior, the posterior reduces to

$$\theta_n = \frac{1}{\sigma^2} V_N X^T y = \frac{1}{\sigma^2}\left(\frac{1}{\tau_0^2} I_d + \frac{1}{\sigma^2} X^T X\right)^{-1} X^T y$$
$$= (\lambda I_d + X^T X)^{-1} X^T y$$

where $\lambda = \sigma^2/\tau_0^2$. We just recovered ridge regression

Also if you make your prior flat, you get maximum likelihood estimate.

Aside.

Let set $\theta_n^T V_n^{-1} - \frac{y^T X}{\sigma^2} - \theta_0^T V_0^{-1} = 0$ and solve for θ_n. This yields

$$\theta_n = V_n \left[V_0^{-1}\theta_0 + \frac{X^T Y}{\sigma^2}\right]$$

And ~~we~~ when this happens, we get

$$P(\theta | X, Y, \sigma^2) \propto e^{-\frac{1}{2}(\theta-\theta_n) V_n^{-1} (\theta - \theta_n)}$$

By the definition of multivariate Gaussian, we have

$$\int e^{-\frac{1}{2}(\theta-\theta_n)^T V_n^{-1}(\theta-\theta_n)} d\theta = |2\pi V_n|^{1/2}$$

$$\therefore \quad P(\theta | X, Y, \sigma^2) = |2\pi V_n|^{-1/2} e^{-1/2(\theta-\theta_n)^T V_n^{-1}(\theta-\theta_n)}$$

You can easily derive this integral from first principles.

* **Q.** So what happens if we have a prior that is not amenable to conjugate analysis? Say I do not know the prior and I picked a uniform distribution.

A. If we donot know the shape of the posterior, we will have to use numerical techniques for ~~evaluat~~ evaluating the integral to find the posterior. We will use for example Monte Carlo techniques.

In this derivation of prior we assumed σ^2 is known. We can also think of a case where a prior on variance is given. There for conjugate analysis we'll use the invers Wishart distribution.

* ML: A probabilistic perspective. Ch. 5

A Theorem for Gaussians (Kevin Murphy's Book)

$$p(x) = \mathcal{N}(x \mid \mu_x, \Sigma_x) \quad \text{margin}$$
$$p(y \mid x) = \mathcal{N}(y \mid Ax + b, \Sigma_y) \quad \text{likelihood} \qquad ①$$

$$p(x \mid y) = \mathcal{N}(x \mid \mu_{x|y}, \Sigma_{x|y})$$
$$\Sigma_{x|y}^{-1} = \Sigma_x^{-1} + A^T \Sigma_y^{-1} A \qquad ②$$
$$\mu_{x|y} = \Sigma_{x|y}\left[A^T \Sigma_y^{-1}(y-b) + \Sigma_x^{-1}\mu_x\right]$$

$$p(y) = \mathcal{N}(y \mid A\mu_x + b, \Sigma_y + A\Sigma_x A^T) \qquad ③$$

The above theorem holds for any two variables x and y. When we were doing the completing squares exercise, I was actually trying to prove box ②.

Recall that we need ①, ② and ③ above to apply the Bayes Rule.

$$p(x \mid y) = \frac{p(y \mid x)\, p(x)}{p(y)}$$

~~Aside~~: the ③ box is often called "convolution". That's because ~~often~~ $p(y) = \int_{x'} p(y \mid x')\, p(x')\, dx'$.

chapter 4 of kevin's book.

Bayesian vs. ML plugin prediction

Posterior mean: $\theta_n = (\lambda I_d + X^T X)^{-1} X^T y$

Posterior variance: $V_n = \sigma^2 (\lambda I_d + X^T X)^{-1}$

To predict, Bayesians marginalize over the posterior: Let x^* be a new input. Then the prediction, given the training data $D = (X, y)$ is:

$$P(y|x^*, D, \sigma^2) = \int \mathcal{N}(y|x^{*T}\theta, \sigma^2)\, \mathcal{N}(\theta|\theta_n, V_n)\, d\theta \quad \text{— (1)}$$

$$= \mathcal{N}(y|x^{*T}\theta_n, \sigma^2 + x_*^T V_n x_*)$$

Or, for each possible value of θ, the prediction is weighted by the posterior. So it is a weighted prediction. It is weighted over an infinite domain. The frequentists to make the prediction use the likelihood.

* Each θ gets weighted by its posterior probability.

This is an example of an <u>ensemble</u> predictor. In contrast an ML predictor is:

$$P(y|x^*, D, \sigma^2) = \mathcal{N}(y|x^{*T}\theta_{ML}, \sigma^2)$$

this assumes that there is one θ.

ML simply computes (re-writes) (1) above as follows:

$$\int \mathcal{N}(y|x^{*T}\theta, \sigma^2)\, \mathcal{N}(\theta|\theta_n, V_n)\, d\theta$$

$$= \int \mathcal{N}(y|x^{*T}\theta, \sigma^2)\, \delta_{\theta_{ML}}(\theta)\, d\theta$$

Delta fn. is just a spike at one place. Also called Dirac fn. or Impulse fn. Integral w.r.t. Delta picks θ_{ML}.

* ML assumes there is only one ~~theta~~ θ.

Bayesians assume there are infinite θs.

Having said above θ_n and θ_{ML} might be quite similar barring any numerical ~~[illegible]~~ issues. However, the term $\cancel{\sigma^2 + x^{*T} V_n x^*}$ $x^{*T} V_n x^*$ is important, which only appears in Bayesian prediction. This term gives us confidence intervals.