Introduction to Probability, Likelihood and Inference

Akshay Badola
15MCPC15

School of Computer and Information Sciences
University of Hyderabad

May, 2018

Overview

Probability

What is Probability?

- "PROBABILITY DOES NOT EXIST" [1]
- "Probabilistic reasoning–always to be understood as subjective–merely stems from our being uncertain about something." [2]

[1] Bruno De Finetti. *Theory of probability: a critical introductory treatment*. Vol. 6. John Wiley & Sons, 2017.

Probability

What is Probability?

- "PROBABILITY DOES NOT EXIST"[1]
- "Probabilistic reasoning–always to be understood as subjective–merely stems from our being uncertain about something." [2]
- Probability mass and measure:
  A function p that assigns to each element x_i of a set X a unique number, such that
  $p(x_i) \geq 0$ and $\sum p(x_i)$ (or $\int p(x)) = 1, \forall x_i \in X$
- The notion of a measure implies when it is alongside the notion of continuum.
- So, for continuous sets, we'll have measure instead of mass.
- Density and Distribution:
  p then is the mass (density) function $P(z)$ is the *distribution* function
  $P(z) = \int_{-\infty}^{z} p(x)dx$, which measures the cumulative probability from $-\infty$ to z then, $p(x) = \frac{dP}{dx}$

[1] Bruno De Finetti. *Theory of probability: a critical introductory treatment*. Vol. 6. John Wiley & Sons, 2017.

Probability

Statistics

- A statistic is any function of data.
- $P(x)$ is then the *distribution* of data. $E(x)$ is the expectation of data.
- $E(x) = \sum xp(x) = \int xp(x)dx$, is the cumulation of the data point multiplied by the probability at that data point.
- $E(x, y) = \sum\sum xyp(x, y) = \int\int xyp(x, y)dxdy$
- If, $P(x, y) = \int\int p(x, y)dxdy$, then $\int p(x, y)dx = ?$

Probability

Statistics

- A statistic is any function of data.
- $P(x)$ is then the *distribution* of data. $E(x)$ is the expectation of data.
- $E(x) = \sum xp(x) = \int xp(x)dx$, is the cumulation of the data point multiplied by the probability at that data point.
- $E(x, y) = \sum\sum xyp(x, y) = \int\int xyp(x, y)dxdy$
- If, $P(x, y) = \int\int p(x, y)dxdy$, then $\int p(x, y)dx =?$

  $= p(y)$
- We can denote $E_{x,y}(p)$, where $p = p(x, y)$ as the expectation w.r.t. to both x and y.
- If we wish to take expectation w.r.t. one variable we write $E_x(p)$.

Some simple things

Some identities

- $p(x, y) = p(x)p(y)$, iff x and y are independent.
- $E(X + Y) = E(X) + E(Y)$: Linearity of expectation.

Some simple things

Some identities

- $p(x, y) = p(x)p(y)$, iff x and y are independent.
- $E(X + Y) = E(X) + E(Y)$: Linearity of expectation.

Given the above prove the following.

- $E(XY) = E(X)E(Y)$ iff X and Y are independent.
- Let $X_1, X_2, ...X_n$ be a sequence of mutually independent and identically distributed random variables such that, Then let $S_k = X_1 + X_2 + ... + X_k$, $1 \leq k \leq n$.
  For $1 \leq m \leq n$, Prove that $E(\frac{S_m}{S_n}) = \frac{m}{n}$
- $E(XY)^2 \leq E(X^2)E(Y^2)$ (Cauchy-Schwarz inequality)

Conditional Probability and Bayes Theorem

Some more simple things

- $p(xy) = p(x|y)p(y) = p(y|x)p(x)$
- $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$. Bayes Rule.

Conditional Probability and Bayes Theorem

Some more simple things

- $p(xy) = p(x|y)p(y) = p(y|x)p(x)$
- $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$. Bayes Rule.
- Similarly we can define
  $E(X|Y) = E(X|Y = y_j) = \frac{\sum x_i P(X{=}x_i;Y{=}y_j)}{P(Y{=}y_j)}$
- Variance is $Var(X) = E[(X - \bar{X})^2]$. Covariance is $Cov(X, Y) = E[(X - \bar{X})(Y - \bar{Y})]$.
- Conditional variance can be similarly defined:
  $Var(X|Y = y_j) = E[X - E(X|Y = y_j)^2|Y = y_j]$

Conditional Probability and Bayes Theorem

Some more simple things

- $p(xy) = p(x|y)p(y) = p(y|x)p(x)$
- $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$. Bayes Rule.
- Similarly we can define
  $E(X|Y) = E(X|Y = y_j) = \frac{\sum x_i P(X=x_i;Y=y_j)}{P(Y=y_j)}$
- Variance is $Var(X) = E[(X - \bar{X})^2]$. Covariance is $Cov(X, Y) = E[(X - \bar{X})(Y - \bar{Y})]$.
- Conditional variance can be similarly defined:
  $Var(X|Y = y_j) = E[X - E(X|Y = y_j)^2|Y = y_j]$

Prove:

- $E(X) = E[E(X|Y)]$
- $Var(X) = E[Var(X|Y)] + Var[E(X|Y)]$
- $Cov(X, Y) = 0, \iff X \perp Y$ (if and only if X is indpendent of Y)

Conditional Probability and Bayes Theorem

The Bayesian Way

- Given a set of data $\mathcal{D}$, often we'd like to say certain things about it. E.g., the data is drawn from a particular distribution, or the probability for a particular point of data is so and so.
- Let's say $f : \S \rightarrow R^k$, where R^k represents some property of a data point that we are hoping to estimate or predict and of course k is the dimensionality of that point.
- From Bayes' rule $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$, or for our problem
  $p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p(x)}{p(\mathcal{D})}$.
  $p(x|\mathcal{D}) \propto p(\mathcal{D}|x)p(x)$
  Or the posterior distribution of the data is proportional to the $Likelihood \times Prior$
- E.g., $f : x \rightarrow < C_1, C_2, ..., C_k >$ can be thought of as a *classifer* function, which maps x to a probability vector of k classes.
- Or $f : y \rightarrow x$ for a given set of data $\mathcal{D} \equiv < x, y >^n$. Such a function would be a *regression* function.

A brief note on likelihood

A little history

- Probability in early 20th century was done mostly in the Laplacian way who was the first to rigorously study the principles.
- Come Fisher. He first published "An absolute criterion for fitting frequency curves" [3], at which time he was working on errors for curve fitting and had zeroed in at least squares estimation for curve fitting.
- A lot of this research was necessitated with the recent advances in biology and genetics.
- Assuming a distribution parametrized by θ, the task is to find such a θ so that $\int y(x) - f(x;\theta)^2 dx$ is minimized.
- Fisher observes that for a reparametrization $x = x(z)$, $\int y(x(z)) - f(x(z);\theta)^2 dz$ the θ is not the same as for the first equation.

A brief note on likelihood

More likelihood history

- He thereby instead based on the hypothesis that the trials are independent and from the same distribution.
  $\text{argmax}_\theta \prod(f(x;\theta))$, which in modern notation we can write as
  $\mathcal{L}(\theta; f, x) = \underset{\theta}{\text{argmax}} \prod(f(x;\theta))$
- Trouble occured with another work with whose authors Fisher had earlier collaborated in "A Cooperative Study"[2]
- For the set of the data they had, they showed that Fisher's method implied placing a uniform prior (or ignorance) over the distribution.
- While from their experience they knew that the data would be clustered around some points. So they rejected Fisher's method and said that it has "academic rather than practical value" saying
  "Statistical workers cannot be too often reminded that there is no validity in a mathematical theory pure and simple. Bayes' theorem must be based on experience..."

[2] John Aldrich. "RA Fisher and the making of maximum likelihood 1912-1922". In: *Statistical science* (1997), pp. 162–176.

Likelihood

Yet more likelihood history

- Then in 1921 and 1922, completely rejected his perceived detractors' reasoning and mentioned he was talking about something else entirely. "Bayes (1763) attempted to find, by observing a sample, the actual probability that the population value p lay in any given range.... Such a problem is indeterminate without knowing the statistical mechanism under which different values of p come into existence; it cannot be solved from the data supplied by a sample, or any number of samples, of the population."
- And thereafter referred to the inference of the parameters via *Likelihood* which arises from some true hidden process and not from data.
- For years Fisher and his followers derided the Bayesian methods, even though the theory was subsequently refined. (In any case they were Laplace's methods)

A brief note on likelihood

From history to present day

- An example of the animosity can be seen in the preface to *A Mathematical Theory of Evidence* "My own work grew out of ...arguments of R. A. Fisher. Beginning about fifty years ago, the British-American school believed that it had, by various pragmatic devices, banished the *Bayesian scourge*..."[3] [emphasis added].
- However, the term was coined so it that when we speak of the parameter θ that maximizes the value of the function $\prod(f(x;\theta))$, it is referred to as *Likelihood*.

[3] Glenn Shafer. *A mathematical theory of evidence*. Vol. 42. Princeton university press, 1976.

Maximum Likelihood Estimation

But first, what is estimation?

- Assuming we have a random sample $x_1, x_2, ...x_n$ drawn independently from some population with a given distribution, which we know (or assume) an estimator $t_n = t(x_1, x_2, ...x_n)$ Is unbiased if $E(t_n) = \theta$.

Prove:

- For a sample from the Gaussian distribution $E(\bar{x}) = \mu$ and $E(S^2) = \sigma^2$ where $S^2 = \frac{1}{n-1}\sum_{1}^{n}(x_i - \bar{x})^2$

Continuing

- According to Fisher, an estimator is *best* if, it is (i) unbiased (ii) consistent (iii) efficient (iv) sufficient
- The quantity $E(t_n) - \theta$ is the bias of the estimator.
- An unbiased estimator $t_n = t(x_1, x_2, ...x_n)$ is called consistent, if for a population parameter θ it converges in probability to true θ as n tends to infinity.
  $P[|t_n - \theta| < \epsilon] > 1 - \eta, \forall\ n > N$
  Or, $P\{\lim_{n\to\infty} = \theta\} = 1 \implies t_n - \frac{P}{inf} > \theta$

More on estimators

Effciency and Sufficiency

- For two consistent estimators t_n and $\acute{t}_n$, t_n is said to be more efficient than $\acute{t}_n$ if $\sigma^2 < \acute{\sigma}^2$.
- An estimator t_n is sufficient for estimating a population parameter θ if knowledge of that estimator alone is enough to obtain that population parameter.
  In terms of the *likelihood function*, if $f(x, \theta)$ is the density for a population,
  then, the likelihood $\mathcal{L}(x_1, x_2, ...x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta)$ with iid assumption.
  If $\mathcal{L}(x_1, x_2, ...x_n; \theta) = \mathcal{L}_1(t_n, \theta)\mathcal{L}_2(x_1, x_2, ...x_n)$
  that is if it can be factorized in terms of a (*likelihood*) function of the parameter and the data separately, then there's no other estimator can provide more information about the data.

Back to Maximum Likelihood

Properties of Maximum Likelihood Estimators

- Maximum likelihood estimators are *biased* (to what?).

Back to Maximum Likelihood

Properties of Maximum Likelihood Estimators

- Maximum likelihood estimators are *biased* (to what?).
- But maximum likelihood estimators (MLE) are consistent.

Back to Maximum Likelihood

Properties of Maximum Likelihood Estimators

- Maximum likelihood estimators are *biased* (to what?).
- But maximum likelihood estimators (MLE) are consistent.
- The distribution of MLEs tends to normal for large samples. Assuming $f(x, \theta)$ is continuous and $\frac{\partial f}{\partial \theta}$ is continuous in an interval containing true θ (say θ_0) and does not vanish, then for large n,

$$\frac{1}{\text{var}\hat{\theta}} = \int_{-\infty}^{\infty} \frac{1}{f}(\frac{\partial f}{\partial \theta})^2 dx = n(\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} \log f)^2 dx$$

$$= nE(\frac{\partial}{\partial \theta} \log f)^2$$

Back to Maximum Likelihood

Properties of Maximum Likelihood Estimators

- Maximum likelihood estimators are *biased* (to what?).
- But maximum likelihood estimators (MLE) are consistent.
- The distribution of MLEs tends to normal for large samples. Assuming $f(x, \theta)$ is continuous and $\frac{\partial f}{\partial \theta}$ is continuous in an interval containing true θ (say θ_0) and does not vanish, then for large n,

$$\frac{1}{\text{var}\hat{\theta}} = \int_{-\infty}^{\infty} \frac{1}{f}(\frac{\partial f}{\partial \theta})^2 dx = n(\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} \log f)^2 dx$$

$$= nE(\frac{\partial}{\partial \theta} \log f)^2$$

- MLEs are most efficient for paramters which are distributed normally about the true parameters.

Back to Maximum Likelihood

Properties of Maximum Likelihood Estimators

- Maximum likelihood estimators are *biased* (to what?).
- But maximum likelihood estimators (MLE) are consistent.
- The distribution of MLEs tends to normal for large samples. Assuming $f(x, \theta)$ is continuous and $\frac{\partial f}{\partial \theta}$ is continuous in an interval containing true θ (say θ_0) and does not vanish, then for large n,

$$\frac{1}{\text{var}\hat{\theta}} = \int_{-\infty}^{\infty} \frac{1}{f}(\frac{\partial f}{\partial \theta})^2 dx = n(\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} \log f)^2 dx$$

$$= nE(\frac{\partial}{\partial \theta} \log f)^2$$

- MLEs are most efficient for paramters which are distributed normally about the true parameters.
- MLEs are sufficient, if sufficient estimators exist.

Back to Maximum Likelihood

Properties of Maximum Likelihood Estimators

- Maximum likelihood estimators are *biased* (to what?).
- But maximum likelihood estimators (MLE) are consistent.
- The distribution of MLEs tends to normal for large samples. Assuming $f(x, \theta)$ is continuous and $\frac{\partial f}{\partial \theta}$ is continuous in an interval containing true θ (say θ_0) and does not vanish, then for large n,

$$\frac{1}{\text{var}\hat{\theta}} = \int_{-\infty}^{\infty} \frac{1}{f}(\frac{\partial f}{\partial \theta})^2 dx = n(\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} \log f)^2 dx$$

$$= nE(\frac{\partial}{\partial \theta} \log f)^2$$

- MLEs are most efficient for paramters which are distributed normally about the true parameters.
- MLEs are sufficient, if sufficient estimators exist.
- MLEs are invariant to reparametrization, i.e., if $\hat{\theta}$ is MLE for θ, then $f(\hat{\theta})$ is MLE for $f(\theta)$.

The Bayesian Way

Accounting for uncertainty

- Even though MLE is consistent in the asymptotic limit, we never really have infinite data on our hands.
- For finite data, as we've seen earlier, it is always biased, which adds uncertainty to the estimated parameter.
- Unfortunately Fisher's methods don't really account for that uncertainty, and we have to go to Bayesian Methods to see why.
- Assuming we have a set of samples from a population drawn indpendently but which are not really representative of the population.
  E.g., we take 100 samples of the heights of school children of a certain grade in a district and assuming a Gaussian Distribution, find the mean and variance as 168cm and 8cm.
- On the other hand, we know that the national statistics for the same population is 174cm and 6cm.
- How do we account for the difference in the data? Do we take the data as it is, or should we account for the uncertainty of the fact that we've taken only 100 samples?

The Bayesian Way

Accounting for uncertainty

- Because of the fact that we've seen only 100 students' heights, and cannot be sure that those are indeed the statistics for the height of the students in that district, it makes to sort of *hedge* our bets, and find some sort of middle ground from the national statistics.
- So perhaps we adjust the statistics a small amount proportional to the difference between the two according to the number of the samples.
- But how much and in what manner should we do that?

The Bayesian Way

Accounting for uncertainty

- Because of the fact that we've seen only 100 students' heights, and cannot be sure that those are indeed the statistics for the height of the students in that district, it makes to sort of *hedge* our bets, and find some sort of middle ground from the national statistics.
- So perhaps we adjust the statistics a small amount proportional to the difference between the two according to the number of the samples.
- But how much and in what manner should we do that?
- Enter the prior: Recall, $p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p(x)}{p(\mathcal{D})}$.

$$p(\theta|x_1, x_2..., x_n) = \frac{p(x_1,x_2,...,x_n|\theta)p(\theta)}{p(x_1,x_2,...,x_n)}.$$
$$\propto p(x_1, x_2, ..., x_n|\theta)p(\theta)$$
$$Posterior \propto Likelihood \times Prior$$

The Bayesian Way

Accounting for uncertainty contd.

- Taking the logarithm.
  $\log p(\theta|x_1, x_2..., x_n) \propto \log p(x_1, x_2, ..., x_n|\theta) + \log p(\theta)$
  And, from optimization theory we know that maximizing the logarithm of a function is equivalent to maxizing the function for a given parameter.
- So generally we deal with *log likelihood* and in the Bayesian case, the *log posterior, log likelihood* and *log prior*.

The Bayesian Way

Accounting for uncertainty contd.

- Taking the logarithm.
  $\log p(\theta|x_1, x_2..., x_n) \propto \log p(x_1, x_2, ..., x_n|\theta) + \log p(\theta)$
  And, from optimization theory we know that maximizing the logarithm of a function is equivalent to maxizing the function for a given parameter.
- So generally we deal with *log likelihood* and in the Bayesian case, the *log posterior, log likelihood* and *log prior*.
- So naturally in this case, if we have information about the population, we can incorporate that information into our model. But what if we don't have any information?

The Bayesian Way

Accounting for uncertainty contd.

- Taking the logarithm.
  $\log p(\theta|x_1, x_2..., x_n) \propto \log p(x_1, x_2, ..., x_n|\theta) + \log p(\theta)$
  And, from optimization theory we know that maximizing the logarithm of a function is equivalent to maxizing the function for a given parameter.
- So generally we deal with *log likelihood* and in the Bayesian case, the *log posterior, log likelihood* and *log prior*.
- So naturally in this case, if we have information about the population, we can incorporate that information into our model. But what if we don't have any information? We'd still like to account for that uncertainty.
- So in the preceding case, the mean would be $\hat{\mu} = \frac{a\mu + b(\frac{1}{n}\sum x_i)}{a+b}$ where, $a = 1/\tau^2, b = n/\sigma^2$, σ^2 is the sample variance and τ^2 is the prior variance.

Estimation for the Gaussian

Everything tends to the Gaussian

- $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$
- $E(x) = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)x dx = \mu$
- $E(x^2) = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)x^2 dx = \mu^2 + \sigma^2$
- $\mathsf{var}(x) = E(x^2) - E^2(x) = \sigma^2$

Estimation for the Gaussian

Everything tends to the Gaussian

- $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$
- $E(x) = \int_{-\infty}^{\infty}\mathcal{N}(x|\mu,\sigma^2)xdx = \mu$
- $E(x^2) = \int_{-\infty}^{\infty}\mathcal{N}(x|\mu,\sigma^2)x^2dx = \mu^2+\sigma^2$
- $\text{var}(x) = E(x^2) - E^2(x) = \sigma^2$
- For the multivariate case
  $\mathcal{N}(\mathbf{x}|\mu,\sigma^2) = \frac{1}{(2\pi|\Sigma|^2)^{1/2}}\exp(-\frac{1}{2}(\mathbf{x}-\mu)^{\mathbf{T}}\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu))$
- For an iid sample

  $$x_1, x_2, ..., x_n \backsim \mathcal{N},\ p(\mathbf{X}|\mu,\sigma^2) = \prod_{n=1}^{N}\mathcal{N}(x_n|\mu,\sigma^2)$$

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

Estimation for the Gaussian

- Or $\log p(\mathbf{X}|\mu, \sigma^2) = \log(\frac{1}{(2\pi\sigma^2)^{1/2}}) \sum_{n=1}^{N} -\frac{1}{2}(\frac{(\mathbf{x}-\mu)}{\sigma})^2$
- $\log p(\mathbf{X}|\mu, \sigma^2) \propto -\sum_{n=1}^{N} (\frac{\mathbf{x}-\mu}{\sigma})^2$ or

  $-\log p(\mathbf{X}|\mu, \sigma^2) \propto \sum_{n=1}^{N} (\frac{\mathbf{x}-\mu}{\sigma})^2$, which is known as the Negative Log Likelihood or NLL for short.
- Since NLL $< 0, \forall x, \mu$, $x = \mu$ maximizes the NLL and hence is the MLE estimator for the mean μ.

Estimation for the Gaussian

- Or $\log p(\mathbf{X}|\mu, \sigma^2) = \log(\frac{1}{(2\pi\sigma^2)^{1/2}}) \sum_{n=1}^{N} -\frac{1}{2}(\frac{(\mathbf{x}-\mu)}{\sigma})^2$
- $\log p(\mathbf{X}|\mu, \sigma^2) \propto -\sum_{n=1}^{N} (\frac{\mathbf{x}-\mu}{\sigma})^2$ or

  $-\log p(\mathbf{X}|\mu, \sigma^2) \propto \sum_{n=1}^{N} (\frac{\mathbf{x}-\mu}{\sigma})^2$, which is known as the Negative Log Likelihood or NLL for short.
- Since NLL $< 0, \forall x, \mu$, $x = \mu$ maximizes the NLL and hence is the MLE estimator for the mean μ.
- Note that it is also the LMS (Least Mean Squared) estimator, so for the gaussian case it coincides. It can be shown that all LMS estimation assumes a Gaussian distribution for the data.
- We'll consider a full bayesian treatment later, for both the univariate and multivariate Gaussian.

Prediction

After paramter estimation

- Let's say we have a set of data $\mathbf{X}, \mathbf{t}$ where $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n})^T$ and $\mathbf{t} = (t_1, t_2, ..., t_n)^T$ are the *input* and *target* variables respectively.
- This is a regression task and a simple formulation exists with the LMS.
- The problem is to fit the input to the target distribution via a linear function.
- Fit is $\mathbf{t} = \mathbf{w}^T\mathbf{X}$, and least squares formulation is $f = (\mathbf{t} - \mathbf{w}^T\mathbf{X})^2$

Prediction

After paramter estimation

- Let's say we have a set of data $\mathbf{X}, \mathbf{t}$ where $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n})^T$ and $\mathbf{t} = (t_1, t_2, ..., t_n)^T$ are the *input* and *target* variables respectively.
- This is a regression task and a simple formulation exists with the LMS.
- The problem is to fit the input to the target distribution via a linear function.
- Fit is $\mathbf{t} = \mathbf{w}^T\mathbf{X}$, and least squares formulation is $f = (\mathbf{t} - \mathbf{w}^T\mathbf{X})^2$
  $\frac{\partial f}{\partial \mathbf{w}} = 2\frac{\partial(-\mathbf{w}^T\mathbf{X})}{\partial \mathbf{w}}(\mathbf{t} - \mathbf{w}^T\mathbf{X}) =$

Prediction

After paramter estimation

- Let's say we have a set of data $\mathbf{X}, \mathbf{t}$ where $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n})^T$ and $\mathbf{t} = (t_1, t_2, ..., t_n)^T$ are the *input* and *target* variables respectively.
- This is a regression task and a simple formulation exists with the LMS.
- The problem is to fit the input to the target distribution via a linear function.
- Fit is $\mathbf{t} = \mathbf{w}^T\mathbf{X}$, and least squares formulation is
  $f = (\mathbf{t} - \mathbf{w}^T\mathbf{X})^2$
  $\frac{\partial f}{\partial \mathbf{w}} = 2\frac{\partial(-\mathbf{w}^T\mathbf{X})}{\partial \mathbf{w}}(\mathbf{t} - \mathbf{w}^T\mathbf{X}) = -\mathbf{X}^T(\mathbf{t} - \mathbf{w}^T\mathbf{X})$
  Or, $\mathbf{X}^T\mathbf{t} = \mathbf{X}^T(\mathbf{w}^T\mathbf{X}) \implies \mathbf{w}\mathbf{X}^T\mathbf{X} = \mathbf{X}^T\mathbf{t}$ (Why?)
  $\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$

Prediction

Predictive distribution

- Now assuming the target distribution for is Gaussian, and for a linear fit
  $p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$ and,
  $p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_i|y(x_i, \mathbf{w}), \beta^{-1})$
- $\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta^{-1}) = -\frac{\beta}{2} \sum_{i=1}^{N} \mathcal{N}(t_i - y(x_i, \mathbf{w})^2 + \frac{N}{2} \ln \beta - \frac{N}{2}(2\pi)$
- Let, $\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{N} (t_i - y(x_i, \mathbf{w}))^2$
  and $\frac{1}{\hat{\beta}} = \frac{1}{N} \sum_{i=1}^{N} (t_i - y(x_i, \mathbf{w}))^2$
- Then the predictive distribution for a target t is defined as
  $p(t|x, \hat{\mathbf{w}}, \hat{\beta}) = \mathcal{N}(t|y(x, \hat{\mathbf{w}}), \hat{\beta}^{-1})$

Prediction

Predictive distribution

- Now assuming the target distribution for is Gaussian, and for a linear fit
  $p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$ and,
  $p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_i|y(x_i, \mathbf{w}), \beta^{-1})$
- $\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta^{-1}) = -\frac{\beta}{2} \sum_{i=1}^{N} \mathcal{N}(t_i - y(x_i, \mathbf{w})^2 + \frac{N}{2} \ln \beta - \frac{N}{2}(2\pi)$
- Let, $\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{N} (t_i - y(x_i, \mathbf{w}))^2$
  and $\frac{1}{\hat{\beta}} = \frac{1}{N} \sum_{i=1}^{N} (t_i - y(x_i, \mathbf{w}))^2$
- Then the predictive distribution for a target t is defined as
  $p(t|x, \hat{\mathbf{w}}, \hat{\beta}) = \mathcal{N}(t|y(x, \hat{\mathbf{w}}), \hat{\beta}^{-1})$
- However, we still haven't accounted for the uncertainty in $\mathbf{w}$. So we put a Gaussian prior over it. $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1})$

Prediction

Posterior Predictive Distribution

- So the posterior becomes: $p(\mathbf{w}|x, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$
- Maximizing with respect to this posterior get a Maximum A Posteriori (MAP) solution

$$\frac{\beta}{2}\sum_{i=1}^{N}(t_i - y(x_i, \mathbf{w}))^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

References I

[1] John Aldrich. "RA Fisher and the making of maximum likelihood 1912-1922". In: *Statistical science* (1997), pp. 162–176 (14).

[2] Bruno De Finetti. *Theory of probability: a critical introductory treatment*. Vol. 6. John Wiley & Sons, 2017 (3, 4).

[3] Ronald A Fisher. "On an absolute criterion for fitting frequency curves". In: *Statistical Science* 12.1 (1997), pp. 39–41 (13).

[4] Glenn Shafer. *A mathematical theory of evidence*. Vol. 42. Princeton university press, 1976 (16).