Frequentist, Conditionalist, and Their Relationship with Machine Learning

Lili Mou
moull12@sei.pku.edu.cn

Outline

Outline

Probability

Kolmogorov (1933):

- Nonnegative
$$p(x) \geq 0, \forall x$$
- Normalized
$$\sum_x p(x) = 1$$
- Finitely/countably additive
Let $A_1, A_2, \cdots, A_n$ be disjoint events,
$$p\left(\bigcup A_i\right) = \sum_i p(A_i)$$

Interpretations

- The probability that a toss of a coin gives the head
- The probability that it will rain tomorrow
- The probability that the speed of light lies in 2.9–3.1 $\times 10^8 m/s$

- **Frequentist** The limit of frequency provided that the number of samples goes to infinity (Recall the Law of Large Numbers)

- **Bayesian** The degree of ones subjective belief
  - Is belief necessarily a kind of probability?
  - Is belief admissible in scientific research? or even unavoidable?

Interpretations

- The probability that a toss of a coin gives the head
- The probability that it will rain tomorrow
- The probability that the speed of light lies in 2.9–3.1 $\times 10^8 m/s$

- **Frequentist** The limit of frequency provided that the number of samples goes to infinity (Recall the Law of Large Numbers)

- **Bayesian** The degree of ones subjective belief
  - Is belief necessarily a kind of probability?
  - Is belief admissible in scientific research? or even unavoidable?

"Some even argue that the frequency concept never applies, it being impossible to have an infinite sequence of i.i.d repetitions of any situation, except in a certain imaginary (subjective) sense." [1]

Interpretations

- The probability that a toss of a coin gives the head
- The probability that it will rain tomorrow
- The probability that the speed of light lies in 2.9–3.1 $\times 10^8 m/s$
- **Frequentist** The limit of frequency provided that the number of samples goes to infinity (Recall the Law of Large Numbers)
- **Bayesian** The degree of ones subjective belief
  - Is belief necessarily a kind of probability?
  - Is belief admissible in scientific research? or even unavoidable?

"Some even argue that the frequency concept never applies, it being impossible to have an infinite sequence of i.i.d repetitions of any situation, except in a certain imaginary (subjective) sense." [1]
"The subjectivist states his judgements, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks under the glorious objectivity of science." (Good, 1973; see also [1].)

Bifurcation of the Two Schools

We have the data $\mathcal{D}$, and the model, parametrized by Θ.

$$\mathcal{D} \sim p_{\Theta}(\cdot)$$

What do we take expectation on (for learning, inference, etc)?

Frequentist: $\mathcal{D}$, because there is nothing random about Θ
No random, no cry

Bayesian: Θ, because $\mathcal{D}$ is known
Everything unknown is a random variable.

Bifurcation of the Two Schools

We have the data $\mathcal{D}$, and the model, parametrized by Θ.

$$\mathcal{D} \sim p_\Theta(\cdot)$$

What do we take expectation on (for learning, inference, etc)?

Frequentist: $\mathcal{D}$, because there is nothing random about Θ
No random, no cry

Bayesian: Θ, because $\mathcal{D}$ is known
Everything unknown is a random variable.

Who is right? Sufficiency + Weak condition $\Rightarrow$ Bayesian analysis

Outline

Pathologies

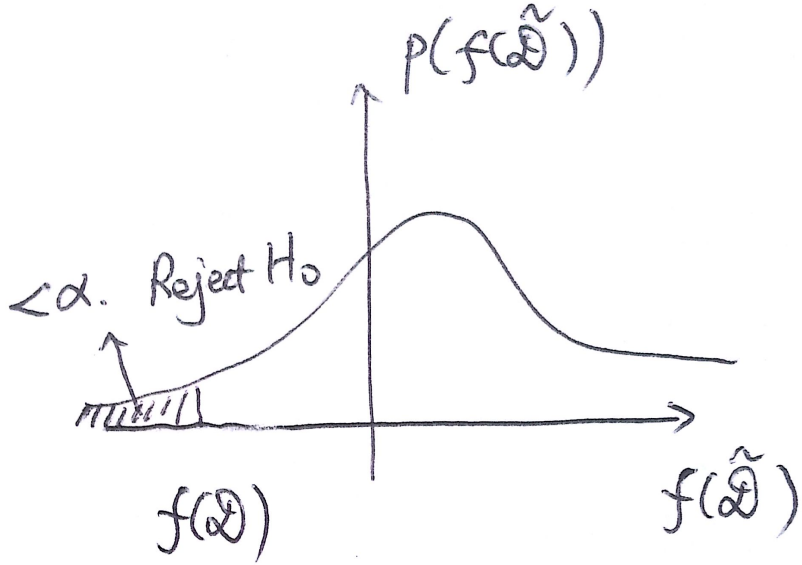
"Frequentist statistics exhibits various forms of weird and undesirable behaviors, known as pathologies." [2]

- Hypothesis test
- Confidence interval

Hypothesis Test and p-Value

- $H_0 \leftrightarrow H_1$
- Data $\mathcal{D}$
- Test statistic: $f(\mathcal{D})$
- p-value$(\mathcal{D}) = \Pr\{f(\tilde{\mathcal{D}}) > f(\mathcal{D}) | \tilde{\mathcal{D}} \sim H_0\}$
- Reject H_0, if p-value$< \alpha$
- Cannot reject H_0, if p-value$\geq \alpha$

Tail Area Probability



First peculiar property: Because p-value relies on the tail area probability, "a hypothesis which may be true may be rejected because it has not predicted observable results which have not occurred." (Jeffreys, 1961; see also [1].)

Who cares about nonoccurrence? But the disaster just begins...

Sample size

Assume $\mathcal{N}(\mu, 1)$

- $H_0 : \mu = 0 \leftrightarrow H_1 : \mu \neq 0$
- True probability: $\mathcal{N}(0.2, 1)$
- $\mathcal{D}_1 = \{0.2\}$, $\mathcal{D}_2$=10000 samples with mean 0.2

Given $\mathcal{D}_1$, we cannot reject H_0.
Given $\mathcal{D}_2$, we do reject H_0.

Sample size

Assume $\mathcal{N}(\mu, 1)$

- $H_0 : \mu = 0 \leftrightarrow H_1 : \mu \neq 0$
- True probability: $\mathcal{N}(0.2, 1)$
- $\mathcal{D}_1 = \{0.2\}$, $\mathcal{D}_2$=10000 samples with mean 0.2

Given $\mathcal{D}_1$, we cannot reject H_0.
Given $\mathcal{D}_2$, we do reject H_0.

For most scientific problems, the only thing that matters is the sample size, which is under control of researchers.

"In a recent survey, 58% of researchers admitted to having collected more data after looking to see whether the results were significant and 22% admitted to stopping an experiment early because they had found the result that they were looking for." (Sanborn et al., 2014)

Interpretation

- Expected type I (false positive) error rate is at most α
- Nothing is said how ofter you errs when you accept, or reject.

Interpretation

- Expected type I (false positive) error rate is at most α
- Nothing is said how ofter you errs when you accept, or reject.

"This is sometimes interpreted as saying that frequentist hypothesis testing is very conservative, since it is unlikely to accidentally reject the null hypothesis. But in fact the opposite is the case: because this method only worries about trying to reject the null, it can never gather evidence in favor of the null, no matter how large the sample size. Because of this, p-values tend to overstate the evidence against the null, and are thus very 'trigger happy.' " [2]

Interpretation

- Expected type I (false positive) error rate is at most α
- Nothing is said how ofter you errs when you accept, or reject.

"This is sometimes interpreted as saying that frequentist hypothesis testing is very conservative, since it is unlikely to accidentally reject the null hypothesis. But in fact the opposite is the case: because this method only worries about trying to reject the null, it can never gather evidence in favor of the null, no matter how large the sample size. Because of this, p-values tend to overstate the evidence against the null, and are thus very 'trigger happy.' " [2]

"It seems somewhat nonsensical, however, that we first deliberately formulate the problem wrong, and then in an *ad hoc* fashion explain the final results in more reasonable terms." [1]

Confidence Interval

$$C_\alpha(\theta) = (l, u) : \Pr\{l(\tilde{\mathcal{D}}) \leq \theta \leq u(\tilde{\mathcal{D}}) | \tilde{\mathcal{D}} \sim \theta\} = 1 - \alpha$$

Counter-intuitive explanation:

- The confidence level (e.g., 95%) is **NOT** the probability that θ lies in the interval, given $\mathcal{D}$.
- It is the probability that the interval covers θ if we repeatedly draw datasets $\tilde{\mathcal{D}}$ (in addition to $\mathcal{D}$ *per se*).
- However, we notice that $\mathcal{D}$ is **KNOWN**. Who cares about nonoccurrence?

Twists and Turns

Pratt (1962):

- An engineer draws a random sample of electron tubes and measures the plate voltages under certain conditions with a very accurate voltmeter, accurate enough so that measurement error is negligible compared with the variability of the tubes.
- A statistician examines the measurements, which look normally distributed and vary from 75 to 99 volts with a mean of 87 and a standard deviation of 4. He makes the ordinary normal analysis, giving a confidence interval for the true mean.

Twists and Turns (2)

- Later he visits the engineer's laboratory, and notices that the voltmeter used reads only as far as 100, so the population appears to be "censored." This necessitates a new analysis, if the statistician is orthodox.
- However, the engineer says he has another meter, equally accurate and reading to 1000 volts, which he would have used if any voltage had been over 100.
- This is a relief to the orthodox statistician, because it means the population was effectively uncensored after all.

Twists and Turns (3)

- But the next day the engineer telephones and says, "I just discovered my high-range voltmeter was not working the day I did the experiment you analyzed for me."
- The statistician ascertains that the engineer would not have held up the experiment until the meter was fixed, and informs him that a new analysis will be required.
- The engineer is astounded. He says, "But the experiment turned out just the same as if the high-range meter had been working. I obtained the precise voltages of my sample anyway, so I learned exactly what I would have learned if the high-range meter had been available. Next you'll be asking about my oscilloscope."

Outline

Weak Conditionality

"Suppose a substance to be analyzed can be sent either to a laboratory in New York or a laboratory in California. The two labs seem equally good, so a fair coin is flipped to choose between them, which "heads" denoting that the lab in New York will be chosen. The coin is flipped and comes up tails, so the California lab is used. After a while, the experimental results come back and a conclusion and report must be developed. Should this conclusion take into account the fact that the coin could have been heads, and hence that the experiment in New York might have been performed instead?"

Common sense (and the conditional viewpoint) cries no, that only the experiment actually performed is relevant, but frequentist reasoning would call for averaging over all possible data, even the possible New York data.

Sufficiency + Weak Conditionality Principle $\Rightarrow$ Bayesian Analysis

Outline

Linear Classification [3]

Assume

$$p(y = i|\boldsymbol{x}) = \sigma(\boldsymbol{w}^T \boldsymbol{x})$$

Given a set of training samples $\boldsymbol{x}_i, y_i$, we would like to predict $y_* =?$ for a new sample $\boldsymbol{x}_*$.

Orthodox Frequentists' Viewpoint

Minimize the expected risk (loss) over $\boldsymbol{x}_*$

$\Downarrow$

Minimize the empirical risk over $\boldsymbol{x}_i, y_i$ (training samples)

$\Downarrow$

Maximize the likelihood of $\boldsymbol{x}_i, y_i$

Maximum likelihood estimation

Training:

$$\boldsymbol{w}^* \leftarrow \arg\max_{\boldsymbol{w}} p(y_i|\boldsymbol{w}, \boldsymbol{x}_i)$$

Predicting:

$$y_* \leftarrow \arg\max_{y} p(y_*|\boldsymbol{x}_*, \boldsymbol{w}^*)$$

Deceitful Frequentists' Viewpoint

We somehow know that God created the world in a neat manner, and thus $\boldsymbol{w}$ is small.

We assume that $\boldsymbol{w} \sim \mathcal{N}(0, \sigma)$ *a priori*.

We therefore maximize the posterior rather than likelihood.

Training:

$$\boldsymbol{w}^* \leftarrow \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|y_i, \boldsymbol{x}_i)$$

$$= \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}) p(y_i|\boldsymbol{w}, \boldsymbol{x}_i)$$

Predicting:

$$y_* \leftarrow \arg\max_{y} p(y_*|\boldsymbol{x}_*; \boldsymbol{w}^*)$$

Gaussian prior $\Rightarrow \ell_2$ penalty
Laplacian prior $\Rightarrow \ell_1$ penalty

Bayesian Learning

- There does not exist $\boldsymbol{w}^*$.
- Rather, $\boldsymbol{w}$ is a random variable that we have to marginalize out.
- Predictive density

$$p(y_*|\boldsymbol{y}) = \int \mathrm{d}\boldsymbol{w}\ p(\boldsymbol{w}|\boldsymbol{y})p(y_*|\boldsymbol{w})$$

Bayesian Learning

- There does not exist $\boldsymbol{w}^*$.
- Rather, $\boldsymbol{w}$ is a random variable that we have to marginalize out.
- Predictive density

$$p(y_*|\boldsymbol{y}) = \int \mathrm{d}\boldsymbol{w}\ p(\boldsymbol{w}|\boldsymbol{y})p(y_*|\boldsymbol{w})$$

Warning: Formula menagerie ahead!

Bayesian Logistic Regression

Let $\boldsymbol{y} \in \{0,1\}^m$ denote the labels of training data $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_m$

Bayesian Logistic Regression

Let $\boldsymbol{y} \in \{0,1\}^m$ denote the labels of training data $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_m$

Prior $p(\boldsymbol{w})$, which is a \$64,000,000 question

Bayesian Logistic Regression

Let $\boldsymbol{y} \in \{0,1\}^m$ denote the labels of training data $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_m$

Prior $p(\boldsymbol{w})$, which is a \$64,000,000 question

Likelihood

$$p(\boldsymbol{y}|\boldsymbol{w}) = \prod_{i=1}^{m} p(y^{(i)}|\boldsymbol{w})$$

Bayesian Logistic Regression

Let $\boldsymbol{y} \in \{0,1\}^m$ denote the labels of training data $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_m$

Prior $p(\boldsymbol{w})$, which is a \$64,000,000 question

Likelihood

$$p(\boldsymbol{y}|\boldsymbol{w}) = \prod_{i=1}^{m} p(y^{(i)}|\boldsymbol{w}) = \prod_{i=1}^{m} \sigma\left(\boldsymbol{w}^T \boldsymbol{\phi}^{(i)}\right)^{t^{(i)}} \left(1 - \sigma\left(\boldsymbol{w}^T \boldsymbol{\phi}^{(i)}\right)\right)^{1-t^{(i)}}$$

Bayesian Logistic Regression

Let $\boldsymbol{y} \in \{0,1\}^m$ denote the labels of training data $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_m$

Prior $p(\boldsymbol{w})$, which is a \$64,000,000 question

Likelihood

$$p(\boldsymbol{y}|\boldsymbol{w}) = \prod_{i=1}^{m} p(y^{(i)}|\boldsymbol{w}) = \prod_{i=1}^{m} \sigma\left(\boldsymbol{w}^T \boldsymbol{\phi}^{(i)}\right)^{t^{(i)}} \left(1 - \sigma\left(\boldsymbol{w}^T \boldsymbol{\phi}^{(i)}\right)\right)^{1-t^{(i)}}$$

Posterior

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{w})p(\boldsymbol{y}|\boldsymbol{w})}{p(\boldsymbol{y})}$$

Bayesian Logistic Regression

Let $\boldsymbol{y} \in \{0,1\}^m$ denote the labels of training data $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_m$

Prior $p(\boldsymbol{w})$, which is a \$64,000,000 question

Likelihood

$$p(\boldsymbol{y}|\boldsymbol{w}) = \prod_{i=1}^{m} p(y^{(i)}|\boldsymbol{w}) = \prod_{i=1}^{m} \sigma\left(\boldsymbol{w}^T \boldsymbol{\phi}^{(i)}\right)^{t^{(i)}} \left(1 - \sigma\left(\boldsymbol{w}^T \boldsymbol{\phi}^{(i)}\right)\right)^{1-t^{(i)}}$$

Posterior

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{w})p(\boldsymbol{y}|\boldsymbol{w})}{p(\boldsymbol{y})} \underset{\boldsymbol{w}}{\propto} p(\boldsymbol{w})p(\boldsymbol{y}|\boldsymbol{w})$$

Bayesian Logistic Regression

Let $\boldsymbol{y} \in \{0,1\}^m$ denote the labels of training data $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_m$

Prior $p(\boldsymbol{w})$, which is a \$64,000,000 question

Likelihood

$$p(\boldsymbol{y}|\boldsymbol{w}) = \prod_{i=1}^{m} p(y^{(i)}|\boldsymbol{w}) = \prod_{i=1}^{m} \sigma\left(\boldsymbol{w}^T \boldsymbol{\phi}^{(i)}\right)^{t^{(i)}} \left(1 - \sigma\left(\boldsymbol{w}^T \boldsymbol{\phi}^{(i)}\right)\right)^{1-t^{(i)}}$$

Posterior

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{w})p(\boldsymbol{y}|\boldsymbol{w})}{p(\boldsymbol{y})} \underset{\boldsymbol{w}}{\propto} p(\boldsymbol{w})p(\boldsymbol{y}|\boldsymbol{w}) = p(\boldsymbol{w}) \prod_{i=1}^{m} \sigma(\cdot)^{t^{(i)}} (1-\sigma(\cdot))^{1-t^{(i)}}$$

Bayesian Logistic Regression

Let $\boldsymbol{y} \in \{0,1\}^m$ denote the labels of training data $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_m$

Prior $p(\boldsymbol{w})$, which is a \$64,000,000 question

Likelihood

$$p(\boldsymbol{y}|\boldsymbol{w}) = \prod_{i=1}^{m} p(y^{(i)}|\boldsymbol{w}) = \prod_{i=1}^{m} \sigma\left(\boldsymbol{w}^T \boldsymbol{\phi}^{(i)}\right)^{t^{(i)}} \left(1 - \sigma\left(\boldsymbol{w}^T \boldsymbol{\phi}^{(i)}\right)\right)^{1-t^{(i)}}$$

Posterior

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{w})p(\boldsymbol{y}|\boldsymbol{w})}{p(\boldsymbol{y})} \underset{\boldsymbol{w}}{\propto} p(\boldsymbol{w})p(\boldsymbol{y}|\boldsymbol{w}) = p(\boldsymbol{w}) \prod_{i=1}^{m} \sigma(\cdot)^{t^{(i)}} (1-\sigma(\cdot))^{1-t^{(i)}}$$

Predictive density

$$p(y_*|\boldsymbol{y}) = \int \mathrm{d}\boldsymbol{w}\ p(y_*|\boldsymbol{w}) \cdot p(\boldsymbol{w}|\boldsymbol{y})$$

$$\underset{y_*}{\propto} \int \mathrm{d}\boldsymbol{w}\ \sigma\left(\boldsymbol{w}^T \boldsymbol{\phi}_*\right) \cdot p(\boldsymbol{w}) \prod_{i=1}^{m} \sigma(\cdot)^{t^{(i)}} (1 - \sigma(\cdot))^{1-t^{(i)}}$$

Intractability

- Posterior is intractable due to the normalizing factor.
- Predictive density is intractable due to the integral.

Intractability

- Posterior is intractable due to the normalizing factor.
- Predictive density is intractable due to the integral.

We have to resort to approximations

- Sampling methods

  Stochastic, usually asymptotically correct, hard to scale
- Deterministic methods

  "Do things wrongly and hope they work"

Laplace Approximation

- Fit a Gaussian at a mode
- The standard deviation is chosen such that ...
  the second-order derivative of the log probability matches
- ☺ The first-order derivative is always 0 at a mode
- ☺ Scale free in representing the unnormalized measure
- ☹ Real variables only
- ☹ Only local properties captured, multi-mode distributions?

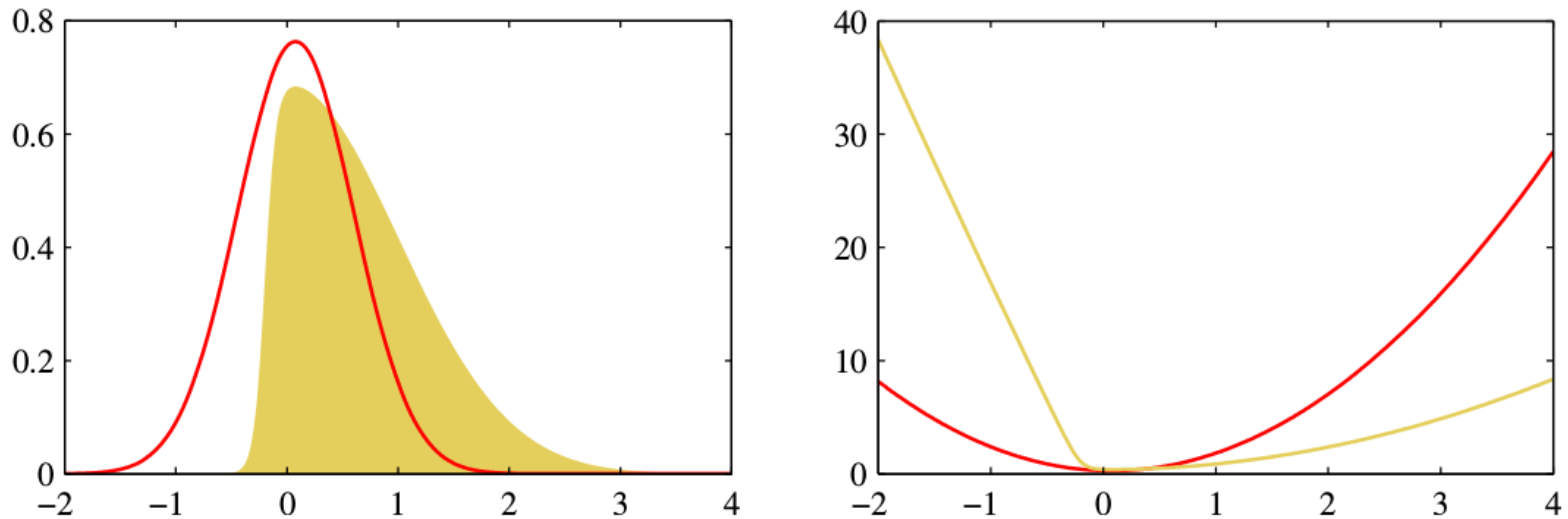


Figure 4.14 Illustration of the Laplace approximation applied to the distribution $p(z) \propto \exp(-z^2/2)\sigma(20z + 4)$ where $\sigma(z)$ is the logistic sigmoid function defined by $\sigma(z) = (1 + e^{-z})^{-1}$. The left plot shows the normalized distribution $p(z)$ in yellow, together with the Laplace approximation centred on the mode z_0 of $p(z)$ in red. The right plot shows the negative logarithms of the corresponding curves.

Fitting a Gaussian

Let $p(z) = \frac{1}{Z} f(z)$ be a true distribution, where $Z = \int f(z)\,\mathrm{d}z$

Step 1: Find a mode z_0 of $p(z)$, by gradient methods, say, satisfying

$$\left.\frac{\mathrm{d}f(z)}{\mathrm{d}z}\right|_{z=z_0} = 0$$

Fitting a Gaussian

Let $p(z) = \frac{1}{Z} f(z)$ be a true distribution, where $Z = \int f(z)\,\mathrm{d}z$

Step 1: Find a mode z_0 of $p(z)$, by gradient methods, say, satisfying

$$\left.\frac{\mathrm{d}f(z)}{\mathrm{d}z}\right|_{z=z_0} = 0$$

Step 2: Consider a Taylor expansion of $\ln f(z)$ at z_0

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A (z - z_0)^2$$

where A is given by

$$A = -\left.\frac{\mathrm{d}^2}{\mathrm{d}z^2} \ln f(z)\right|_{z=z_0}$$

Taking the exponential,

$$f(z) \simeq f(z_0) \exp\left\{-\frac{A}{2}(z - z_0)^2\right\}$$

Fitting a Gaussian

Let $p(z) = \frac{1}{Z} f(z)$ be a true distribution, where $Z = \int f(z)\,\mathrm{d}z$

Step 1: Find a mode z_0 of $p(z)$, by gradient methods, say, satisfying

$$\left.\frac{\mathrm{d}f(z)}{\mathrm{d}z}\right|_{z=z_0} = 0$$

Step 2: Consider a Taylor expansion of $\ln f(z)$ at z_0

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A (z - z_0)^2$$

where A is given by

$$A = -\left.\frac{\mathrm{d}^2}{\mathrm{d}z^2} \ln f(z)\right|_{z=z_0}$$

Taking the exponential,

$$f(z) \simeq f(z_0) \exp\left\{-\frac{A}{2}(z - z_0)^2\right\}$$

Step 3: Normalize to a Gaussian distribution

$$q(z) = \left(\frac{A}{2\pi}\right)^{1/2} \exp\left\{-\frac{A}{2}(z - z_0)^2\right\}$$

Laplace Approximation for Multivariate Distributions

To approximate $p(\boldsymbol{z}) = \frac{1}{Z} f(\boldsymbol{z})$, where $\boldsymbol{z} \in \mathbb{R}^m$

We expand at mode $\boldsymbol{z}_0$

$$\ln f(\boldsymbol{z}) \simeq \ln f(\boldsymbol{z}_0) - \frac{1}{2}(\boldsymbol{z} - z_0)^T \boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0)$$

where $\boldsymbol{A} = -\nabla\nabla \ln f(\boldsymbol{z})\Big|_{\boldsymbol{z}=\boldsymbol{z}_0}$, Hessian of $\ln f(\boldsymbol{z})$, serving as the *precision matrix* in a Gaussian distribution

Taking the exponential, we obtain

$$f(\boldsymbol{z}) \simeq f(\boldsymbol{z}_0) \exp\left\{-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)^T \boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0)\right\}$$

Normalize it as a distribution, and then we have

$$q(\boldsymbol{z}) = \frac{|\boldsymbol{A}|^{1/2}}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)^T \boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0)\right\} = \mathcal{N}(\boldsymbol{z}|\boldsymbol{z}_0, \boldsymbol{A}^{-1})$$

Bayesian Logistic Regression: A Revisit in Earnest

Prior: Gaussian, which is natural[1]

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)$$

Posterior:
$$p(\boldsymbol{w}|\boldsymbol{t}) \propto p(\boldsymbol{w})p(\boldsymbol{t}|\boldsymbol{w})$$

$$\ln p(\boldsymbol{w}|\boldsymbol{t}) = -\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0) \quad \text{[prior]}$$

$$+\sum_{i=1}^{m}\left\{t^{(i)}\ln y^{(i)} + \left(1-t^{(i)}\right)\ln\left(1-y^{(i)}\right)\right\} \quad \text{[likelihood]}$$

$$+ \text{const}$$

$$\boldsymbol{S}_N = -\nabla\nabla \ln p(\boldsymbol{w}|\boldsymbol{t}) = \boldsymbol{S}_0^{-1} + \sum_{i=1}^{m} y^{(i)}(1-y^{(i)})\phi_n\phi_n^T$$

Hence, the Laplace approximation to the posterior is

$$q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_{\text{MAP}}, \boldsymbol{S}_N)$$

[1] Mathematicians always choose priors for the sake of convenience rather than approaching God.

Predictive Density

$$p(\mathcal{C}_1|\boldsymbol{\phi}_*, \boldsymbol{t}) = \int p(\mathcal{C}_1|\boldsymbol{\phi}_*, \boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{t}) \, \mathrm{d}\boldsymbol{w} \simeq \int \sigma(\boldsymbol{w}^T \boldsymbol{\phi}_*) q(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w}$$

$$p(\mathcal{C}_2|\boldsymbol{\phi}_*, \boldsymbol{t}) = 1 - p(\mathcal{C}_1|\boldsymbol{\phi}_*, \boldsymbol{t})$$

Plan

- Change it to a univariate integral
- Substitute sigmoid with a probit function, which is then convolved with a normal

$$\int \sigma(\cdot) \mathcal{N}(\cdot) \, \mathrm{d}\cdot \simeq \int \Phi(\cdot) \mathcal{N}(\cdot) \, \mathrm{d}\cdot = \Phi(\cdot) \simeq \sigma(\cdot)$$

The Dirac Delta Fucntion

Let δ be the *Dirac delta* function, loosely thought of a function such that

- Gaussian distribution peaked at 0 with standard deviation $\to 0$
- $\delta(x) = \begin{cases} +\infty, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases}$

δ function satisfies

$$\int \delta(a - x) f(a) \, \mathrm{d}a = f(x)$$

and specifically

$$\int \delta(a - x) \, \mathrm{d}a = 1$$

Deriving the Predictive Density

$$p(\mathcal{C}_1|\boldsymbol{\phi}_*) \simeq \int \sigma(\boldsymbol{w}^T\boldsymbol{\phi}_*)q(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} \qquad \text{[Laplace approx.]}$$

$$\sigma(\boldsymbol{w}^T\boldsymbol{\phi}) = \int \delta(a - \boldsymbol{w}^T\boldsymbol{\phi}_*)\sigma(a)\,\mathrm{d}a \qquad \text{[Def. of } \delta\text{]}$$

$$\begin{aligned} p(\mathcal{C}_1|\boldsymbol{\phi}_*) &= \int\int \delta(a - \boldsymbol{w}^T\boldsymbol{\phi})\sigma(a)\,\mathrm{d}a\, q(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w} \\ &= \int\int \sigma(a)\delta(a - \boldsymbol{w}^T\boldsymbol{\phi}_*)q(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}\,\mathrm{d}a \\ &\triangleq \int \sigma(a)p(a)\,\mathrm{d}a \end{aligned}$$

where

$$p(a) \triangleq \int \delta(a - \boldsymbol{w}^T\boldsymbol{\phi}_*)q(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}$$

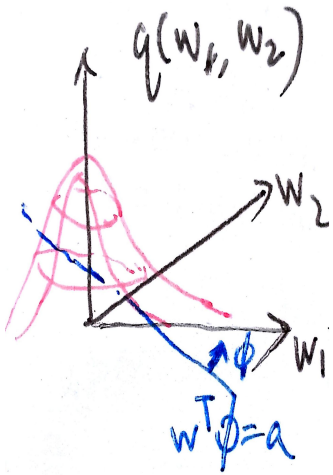
We now argue that $p(a)$ is Gaussian

Deriving the Predictive Density (2)

$$p(a) = \int\int\cdots\int \delta(a - \boldsymbol{w}^T\boldsymbol{\phi}_*)q(\boldsymbol{w})\,\mathrm{d}w_1\,\mathrm{d}w_2\cdots\mathrm{d}w_n$$

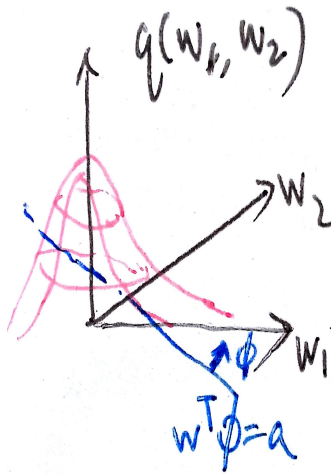
$$= \int\int\cdots\int q(\tilde{\boldsymbol{w}})\,\mathrm{d}w_2\cdots\mathrm{d}w_n$$

$\tilde{\boldsymbol{w}}$ is such that $\tilde{\boldsymbol{w}}^T\boldsymbol{\phi}_* = a$



Deriving the Predictive Density (2)

$$p(a) = \int\int\cdots\int \delta(a - \boldsymbol{w}^T\boldsymbol{\phi}_*)q(\boldsymbol{w})\,\mathrm{d}w_1\,\mathrm{d}w_2\cdots\mathrm{d}w_n$$
$$= \int\int\cdots\int q(\tilde{\boldsymbol{w}})\,\mathrm{d}w_2\cdots\mathrm{d}w_n$$



$\tilde{\boldsymbol{w}}$ is such that $\tilde{\boldsymbol{w}}^T\boldsymbol{\phi}_* = a$

We can also verify that

$$\int p(a)\,\mathrm{d}a = \int\int \delta(a - \boldsymbol{w}^T\boldsymbol{\phi}_*)q(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}\,\mathrm{d}a$$
$$= \int\int \delta(a - \boldsymbol{w}^T\boldsymbol{\phi}_*)q(\boldsymbol{w})\,\mathrm{d}a\,\mathrm{d}\boldsymbol{w}$$
$$= \int q(\boldsymbol{w})\,\mathrm{d}\,\boldsymbol{w}\int \delta(a - \boldsymbol{w}^T\boldsymbol{\phi})\,\mathrm{d}a$$
$$= 1$$

Deriving the Predictive Density (3)

$$\mu_a = \mathbb{E}[a] = \int p(a)a \, \mathrm{d}a = \int q(\boldsymbol{w})\boldsymbol{w}^T\boldsymbol{\phi}_* \, \mathrm{d}\boldsymbol{w} = \boldsymbol{w}_{\mathsf{MAP}}^T\boldsymbol{\phi}_*$$

$$\begin{aligned}\sigma_a^2 = \mathsf{var}[a] &= \int p(a)\left\{a^2 - \mathbb{E}[a]^2\right\} \mathrm{d}a \\ &= \int q(\boldsymbol{w})\left\{(\boldsymbol{w}^T\boldsymbol{\phi})^2 - (\boldsymbol{m}_N^2\boldsymbol{\phi})^2\right\} \mathrm{d}\boldsymbol{w} \\ &= \phi^T\boldsymbol{S}_N\boldsymbol{\phi}\end{aligned}$$

Thus

$$\begin{aligned}p(\mathcal{C}_1|\boldsymbol{t}) &\simeq \int \sigma(a)p(a) \, \mathrm{d}a = \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2) \, \mathrm{d}a \\ &\simeq \int \Phi(\lambda a)\mathcal{N}(a|\mu_a, \sigma_a^2) \, \mathrm{d}a = \Phi\left(\frac{\mu_a}{(\lambda^{-2} + \sigma_a^2)^{1/2}}\right) \simeq \sigma\left(\kappa(\sigma_a^2)\mu_a\right)\end{aligned}$$

$\lambda = \sqrt{\pi/8}$, $\kappa(\sigma_a^2) = (1 + \pi\sigma_a^2/8)^{-1/2}$, chosen such that the rescaled probit function has the same slope as sigmoid at the origin.

Hiearchical Bayes

What if we have parameters in the prior?

- Maximum likelihood estimation (Empirical Bayes, Type-II ML)
- Max *a posteriori*, assuming some prior on the hyper-parameters
- Full Bayesian treatment: Marginalize out all unknown variable!

Take-Home Messages

- Frequentist takes expectation on (known) data while conditioning on (unknown) θ
- Conditionalist (Bayesian) takes expectation on (unknown) θ while conditioning on (known) data
- Bayes Hierarchy
  - MLE
  - Max *a posteriori*
  - Empirical Bayes
  - Max *a posteriori* estimation of hyper-parameters
  - Full Bayes
- Bayesian treatment is fundamentally correct by computationally non-trivial.

References

[1] James O. Berger, *Statistical Decision Theory and Bayesian Analysis (2 ed.)*, Springer-Verlag, 1985.

[2] Kevin P. Murphy, *Machine Learning: A Probability Perspective*, MIT Press, 2012.

[3] Christopher M. Bishop *Pattern Recognition and Machine Learning*, Springer, 2006.