0 Statistical Learning Basics

Goal

Maximum Likelihood, Prior, Posterior, MAP, Bayesian LR

Alert 0.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.
This note is likely to be updated again soon.

Definition 0.2: Distribution and density

Recall that the cumulative distribution function (cdf) of a random vector $\mathbf{X} \in \mathbb{R}^d$ is defined as:

$$F(\mathbf{x}) := \Pr(\mathbf{X} \leq \mathbf{x}),$$

and its probability density function (pdf) is

$$p(\mathbf{x}) := \frac{\partial^d F}{\partial x_1 \cdots \partial x_d}(\mathbf{x}), \text{ or equivalently } F(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} p(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

Clearly, each cdf $F : \mathbb{R}^d \to [0, 1]$ is

- monotonically increasing in each of its inputs;
- right continuous in each of its inputs;
- $\lim_{\mathbf{x}\to\infty} F(\mathbf{x}) = 1$ and $\lim_{\mathbf{x}\to-\infty} F(\mathbf{x}) = 0$.

On the other hand, each pdf $p : \mathbb{R}^d \to \mathbb{R}_+$

- integrates to 1, i.e. $\int_{-\infty}^{\infty} p(\mathbf{x})\, \mathrm{d}\mathbf{x} = 1$.

(The cdf and pdf of a discrete random variable can be defined similarly and is omitted.)

Remark 0.3: Change-of-variable

Let $\mathsf{T} : \mathbb{R}^d \to \mathbb{R}^d$ be a diffeomorphism (differentiable bijection with differentiable inverse). Let $\mathbf{X} = \mathsf{T}(\mathbf{Z})$, then we have the change-of-variable formula for the pdfs:

$$p(\mathbf{x})\, \mathrm{d}\mathbf{x} \approx q(\mathbf{z})\, \mathrm{d}\mathbf{z}, \text{ i.e. } p(\mathbf{x}) = q(\mathsf{T}^{-1}(\mathbf{x})) \left| \det \frac{\mathrm{d}\mathsf{T}^{-1}}{\mathrm{d}\mathbf{x}}(\mathbf{x}) \right|$$
$$q(\mathbf{z}) = p(\mathsf{T}(\mathbf{z})) \left| \det \frac{\mathrm{d}\mathsf{T}}{\mathrm{d}\mathbf{z}}(\mathbf{z}) \right|,$$

where det denotes the determinant.

Definition 0.4: Marginal, conditional, and independence

Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ be a random vector with pdf $p(\mathbf{x}) = p(\mathbf{x}_1, \mathbf{x}_2)$. We say $\mathbf{X}_1$ is a marginal of $\mathbf{X}$ with pdf

$$p_1(\mathbf{x}_1) = \int_{-\infty}^{\infty} p(\mathbf{x}_1, \mathbf{x}_2)\, \mathrm{d}\mathbf{x}_2,$$

where we marginalize over $\mathbf{X}_2$ by integrating it out. Similarly $\mathbf{X}_2$ is a marginal of $\mathbf{X}$ with pdf

$$p_2(\mathbf{x}_2) = \int_{-\infty}^{\infty} p(\mathbf{x}_1, \mathbf{x}_2)\,\mathrm{d}\mathbf{x}_1.$$

We then define the conditional $\mathbf{X}_1|\mathbf{X}_2$ with density:

$$p_{1|2}(\mathbf{x}_1|\mathbf{x}_2) = p(\mathbf{x}_1, \mathbf{x}_2)/p_2(\mathbf{x}_2),$$

where the value of $p_{1|2}$ is arbitrary if $p_2(\mathbf{x}_2) = 0$ (usually immaterial). Similarly we may define the conditional $\mathbf{X}_2|\mathbf{X}_1$. It is obvious from our definition that

$$p(\mathbf{x}_1, \mathbf{x}_2) = p_1(\mathbf{x}_1)p_{2|1}(\mathbf{x}_2|\mathbf{x}_1) = p_2(\mathbf{x}_2)p_{1|2}(\mathbf{x}_1|\mathbf{x}_2),$$

namely the joint density p can be factorized into the product of marginal p_1 and conditional $p_{2|1}$. Usually, we omit all subscripts in p when referring to the marginal or conditional whenever the meaning is obvious from context.

Iterating the above construction, we obtain the famous chain rule:

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d) = \prod_{j=1}^{d} p(\mathbf{x}_j|\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}),$$

with obviously $p(\mathbf{x}_1|\mathbf{x}_1, \ldots, \mathbf{x}_0) := p(\mathbf{x}_1)$. We say that the random vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_d$ are independent if

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d) = \prod_{j=1}^{d} p(\mathbf{x}_j).$$

All of our constructions above can be done with cdfs as well (with serious complication for the conditional though). In particular, we have the Bayes rule:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)} = \frac{\Pr(B|A)\Pr(A)}{\Pr(B, A) + \Pr(B, \neg A)}.$$

Definition 0.5: Mean, variance and covariance

Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a random (column) vector. We define its mean (vector) as

$$\boldsymbol{\mu} = \mathbb{E}\mathbf{X}, \quad \text{where} \quad \mu_j = \int x_j \cdot p(x_j)\,\mathrm{d}x_j$$

and its covariance (matrix) as

$$\Sigma = \mathbb{E}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top, \quad \text{where} \quad \Sigma_{ij} = \int (x_i - \mu_i)(x_j - \mu_j) \cdot p(x_i, x_j)\,\mathrm{d}x_i\,\mathrm{d}x_j.$$

By definition Σ is symmetric $\Sigma_{ij} = \Sigma_{ji}$ and positive semidefinite (all eigenvalues are nonnegative). The j-th diagonal entry of the covariance $\sigma_j^2 := \Sigma_{jj}$ is called the variance of X_j.

Exercise 0.6: Covariance

Prove the following equivalent formula for the covariance:

- $\Sigma = \mathbb{E}\mathbf{X}\mathbf{X}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top$;

- $\Sigma = \frac{1}{2}\mathbb{E}(\mathbf{X} - \mathbf{X}')(\mathbf{X} - \mathbf{X}')^\top$, where $\mathbf{X}'$ is iid (independent and identically distributed) with $\mathbf{X}$.

Suppose $\mathbf{X}$ has mean $\boldsymbol{\mu}$ and covariance Σ. Find the mean and covariance of $A\mathbf{X} + \mathbf{b}$, where $A, \mathbf{b}$ are deterministic.

Example 0.7: Multivariate Gaussian

The pdf of the multivariate Gaussian distribution (a.k.a. normal distribution) is:

$$p(\mathbf{x}) = (2\pi)^{-d/2}[\det(\Sigma)]^{-1/2} \exp\left(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right),$$

where d is the dimension and det denotes the determinant of a matrix. We typically use the notation $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = \mathbb{E}\mathbf{X}$ is its mean and $\Sigma = \mathbb{E}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top$ is its covariance.

An important property of the multivariate Gaussian distribution is its equivariance under affine transformations:

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \implies A\mathbf{X} + \mathbf{b} \sim \mathcal{N}(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^\top).$$

(This property actually characterizes the multivariate Gaussian distribution.)

Exercise 0.8: Marginal and conditional of multivariate Gaussian

Let $\begin{bmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)$. Prove the following results:

$$\mathbf{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_{11}), \quad \mathbf{X}_2|\mathbf{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{X}_1 - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12});$$
$$\mathbf{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_{22}), \quad \mathbf{X}_1|\mathbf{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

Remark 0.9: Bias-variance trade-off

Suppose we are interested in predicting a random (scalar) quantity Y based on some feature vector (a.k.a. covariate) $\mathbf{X}$, using the function $\hat{f}$. Here the hat notation suggests $\hat{f}$ may depend on other random quantities, such as samples from a training set. Often, we use the squared loss to evaluate our prediction:

$$\begin{aligned}\mathbb{E}(\hat{f}(\mathbf{X}) - Y)^2 &= \mathbb{E}\Big(\hat{f}(\mathbf{X}) - \mathbb{E}\hat{f}(\mathbf{X}) + \mathbb{E}\hat{f}(\mathbf{X}) - \mathbb{E}(Y|\mathbf{X}) + \mathbb{E}(Y|\mathbf{X}) - Y\Big)^2 \\ &= \underbrace{\mathbb{E}\big(\hat{f}(\mathbf{X}) - \mathbb{E}\hat{f}(\mathbf{X})\big)^2}_{\text{variance}} + \underbrace{\mathbb{E}\big(\mathbb{E}\hat{f}(\mathbf{X}) - \mathbb{E}(Y|\mathbf{X})\big)^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\big(\mathbb{E}(Y|\mathbf{X}) - Y\big)^2}_{\text{difficulty}},\end{aligned}$$

where recall that $\mathbb{E}(Y|\mathbf{X})$ is the so-called regression function. The last term indicates the difficulty of our problem and cannot be reduced by our choice of $\hat{f}$. The first two terms reveals an inherent trade-off in designing $\hat{f}$:

- the variance term reflects the fluctuation incurred by training on some random training set. Typically, a less flexible $\hat{f}$ will incur a smaller variance (e.g. constant functions have 0 variance);
- the (squared) bias term reflects the mismatch of our choice of $\hat{f}$ and the optimal regression function. Typically, a very flexible $\hat{f}$ will incur a smaller bias (e.g. when $\hat{f}$ can model *any* function).

The major goal of much of ML is to strike an appropriate balance between the first two terms.

Definition 0.10: Maximum likelihood estimation(MLE)

Suppose we have a dataset $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, where each sample $\mathbf{x}_i$ (is assumed to) follow some pdf $p(\mathbf{x}|\theta)$ with *unknown* parameter θ. We define the likelihood of a parameter θ given the dataset $\mathcal{D}$ as:

$$L(\theta) = L(\theta; \mathcal{D}) := p(\mathcal{D}|\theta) = \prod_{i=1}^{n} p(\mathbf{x}_i|\theta),$$

where in the last equality we assume our data is iid. A popular way to find an estimate of the parameter θ is to maximize the likelihood over some parameter space Θ:

$$\theta_{\mathsf{MLE}} := \operatorname{argmax}_{\theta\in\Theta} L(\theta).$$

Equivalently, by taking the log and negating, we minimize the negative log-likelihood (NLL):

$$\theta_{\mathsf{MLE}} := \operatorname{argmin}_{\theta\in\Theta} \sum_{i=1}^{n} -\log p(\mathbf{x}_i|\theta).$$

We remark that MLE is applicable only when we can evaluate the likelihood function efficiently, which turns out to be not the case in many settings and we will study alternative algorithms.

Example 0.11: Sample mean and covariance as MLE

Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be iid samples from the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ where the parameters $\boldsymbol{\mu}$ and Σ are to be found. We apply maximum likelihood:

$$\hat{\boldsymbol{\mu}}_{\mathsf{MLE}} := \operatorname*{argmin}_{\boldsymbol{\mu}} \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}).$$

Applying Theorem -1.24 we obtain the sample mean:

$$\hat{\boldsymbol{\mu}}_{\mathsf{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i =: \hat{\mathbb{E}}\mathbf{x},$$

where the hat expectation $\hat{\mathbb{E}}$ is w.r.t. the given data.

Similarly we can show

$$\hat{\Sigma}_{\mathsf{MLE}} := \operatorname*{argmin}_{\Sigma} \log\det\Sigma + \sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}).$$

Or equivalently

$$\hat{\Sigma}^{-1}_{\mathsf{MLE}} := \operatorname*{argmin}_{S} -\log\det S + \sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^\top S(\mathbf{x}_i - \boldsymbol{\mu}).$$

Applying Theorem -1.24 (with the fact that the gradient of $\log\det S$ is S^{-1}), we obtain:

$$\hat{\Sigma}_{\mathsf{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top = \hat{\mathbb{E}}\mathbf{x}\mathbf{x}^\top - (\hat{\mathbb{E}}\mathbf{x})(\hat{\mathbb{E}}\mathbf{x})^\top,$$

where we plug in the ML estimate $\hat{\boldsymbol{\mu}}_{\mathsf{MLE}}$ of $\boldsymbol{\mu}$ if it is not known.

Exercise 0.12: Bias and variance of sample mean and covariance

Calculate the following bias and variance:

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\mathsf{MLE}}] &= \\ \mathbb{E}[\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\mathsf{MLE}}][\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\mathsf{MLE}}]^\top &= \\ \mathbb{E}[\Sigma - \hat{\Sigma}_{\mathsf{MLE}}] &=\end{aligned}$$

Definition 0.13: f-divergence (Csiszár 1963; Ali and Silvey 1966)

Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a strictly convex function (see Definition -1.9) with $f(1) = 0$. We define the following f-divergence to measure the closeness of two pdfs p and q:

$$\mathsf{D}_f(p\|q) := \int f\big(p(\mathbf{x})/q(\mathbf{x})\big) \cdot q(\mathbf{x})\,\mathrm{d}\mathbf{x},$$

where we assume $q(\mathbf{x}) = 0 \implies p(\mathbf{x}) = 0$ (otherwise we put the divergence to ∞).

Csiszár, I. (1963). "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten". *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének közleményei*, vol. 8, pp. 85–108.

Ali, S. M. and S. D. Silvey (1966). "A General Class of Coefficients of Divergence of One Distribution from Another". *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 28, no. 1, pp. 131–142.

Exercise 0.14: Properties of f-divergence

Prove the following:

- $\mathsf{D}_f(p\|q) \geq 0$, with 0 attained iff $p = q$;
- $\mathsf{D}_{f+g} = \mathsf{D}_f + \mathsf{D}_g$ and $\mathsf{D}_{sf} = s\mathsf{D}_f$ for $s > 0$;
- Let $g(t) = f(t) + s(t-1)$ for any s. Then, $\mathsf{D}_g = \mathsf{D}_f$;
- If $p(\mathbf{x} = 0) \iff q(\mathbf{x}) = 0$, then $\mathsf{D}_f(p\|q) = \mathsf{D}_{f^\diamond}(q\|p)$, where $f^\diamond(t) := t \cdot f(1/t)$;
- $f^\diamond$ is (strictly) convex, $f^\diamond(1) = 0$ and $(f^\diamond)^\diamond = f$;

The second last result indicates that f-divergences are not usually symmetric. However, we can always symmetrize them by the transformation: $f \leftarrow f + f^\diamond$.

Example 0.15: KL and LK

Let $f(t) = t \log t$, then we obtain the Kullback-Leibler (KL) divergence:

$$\mathsf{KL}(p\|q) = \int p(\mathbf{x}) \log(p(\mathbf{x})/q(\mathbf{x}))\,\mathrm{d}\mathbf{x}.$$

Reverse the inputs we obtain the reverse KL divergence:

$$\mathsf{LK}(p\|q) := \mathsf{KL}(q\|p).$$

Verify by yourself that the underlying function $f = -\log$ for reverse KL.

Definition 0.16: Entropy, conditional entropy, cross-entropy, and mutual information

We define the entropy of a random vector $\mathbf{X}$ with pdf p as:

$$\mathbb{H}(\mathbf{X}) := \mathbb{E} - \log p(\mathbf{X}) = -\int p(\mathbf{x}) \log p(\mathbf{x}) \, \mathrm{d}\mathbf{x},$$

the conditional entropy between $\mathbf{X}$ and $\mathbf{Z}$ (with pdf q) as:

$$\mathbb{H}(\mathbf{X}|\mathbf{Z}) := \mathbb{E} - \log p(\mathbf{X}|\mathbf{Z}) = -\int p(\mathbf{x}, \mathbf{z}) \log p(\mathbf{x}|\mathbf{z}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{z},$$

and the cross-entropy between $\mathbf{X}$ and $\mathbf{Z}$ as:

$$\mathbb{\dagger}(\mathbf{X}, \mathbf{Z}) := \mathbb{E} - \log q(\mathbf{X}) = -\int p(\mathbf{x}) \log q(\mathbf{x}) \, \mathrm{d}\mathbf{x}.$$

Finally, we define the mutual information between $\mathbf{X}$ and $\mathbf{Z}$ as:

$$\mathbb{I}(\mathbf{X}, \mathbf{Z}) := \mathsf{KL}(p(\mathbf{x}, \mathbf{z}) \| p(\mathbf{x}) q(\mathbf{z})) = \int p(\mathbf{x}, \mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x}) q(\mathbf{z})} \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{z}$$

Exercise 0.17: Information theory

Verify the following:

$$\begin{aligned}
\mathbb{H}(\mathbf{X}, \mathbf{Z}) &= \mathbb{H}(\mathbf{Z}) + \mathbb{H}(\mathbf{X}|\mathbf{Z}) \\
\mathbb{\dagger}(\mathbf{X}, \mathbf{Z}) &= \mathbb{H}(\mathbf{X}) + \mathsf{KL}(\mathbf{X} \| \mathbf{Z}) = \mathbb{H}(\mathbf{X}) + \mathsf{LK}(\mathbf{Z} \| \mathbf{X}) \\
\mathbb{I}(\mathbf{X}, \mathbf{Z}) &= \mathbb{H}(\mathbf{X}) - \mathbb{H}(\mathbf{X}|\mathbf{Z}) \\
\mathbb{I}(\mathbf{X}, \mathbf{Z}) &\geq 0, \text{ with equality iff } \mathbf{X} \text{ independent of } \mathbf{Z} \\
\mathsf{KL}(p(\mathbf{x}, \mathbf{z}) \| q(\mathbf{x}, \mathbf{z})) &= \mathsf{KL}(p(\mathbf{z}) \| q(\mathbf{z})) + \mathbb{E}[\mathsf{KL}(p(\mathbf{x}|\mathbf{z}) \| q(\mathbf{x}|\mathbf{z}))].
\end{aligned}$$

All of the above can obviously be iterated to yield formula for more than two random vectors.

Exercise 0.18: Multivariate Gaussian

Compute

- the entropy of the multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$;
- the KL divergence between two multivariate Gaussians $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ and $\mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$.

Example 0.19: More divergences, more fun

Derive the formula for the following f-divergences:

- χ^2-divergence: $f(t) = (t-1)^2$;
- Hellinger divergence: $f(t) = (\sqrt{t} - 1)^2$;
- total variation: $f(t) = |t - 1|$;
- Jensen-Shannon divergence: $f(t) = t \log t - (t+1) \log(t+1) + \log 4$;
- Rényi divergence (Rényi 1961): $f(t) = \frac{t^\alpha - 1}{\alpha - 1}$ for some $\alpha > 0$ (for $\alpha = 1$ we take limit and obtain ?).

Which of the above are symmetric?

Rényi, A. (1961). "On Measures of Entropy and Information". In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 547–561.

Remark 0.20: MLE = KL minimization

Let us define the empirical "pdf" based on a dataset $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$:

$$\hat{p}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i},$$

where $\delta_{\mathbf{x}}$ is the "illegal" delta mass concentrated at $\mathbf{x}$. Then, we claim that

$$\theta_{\mathsf{MLE}} = \operatorname*{argmin}_{\theta\in\Theta} \ \mathsf{KL}\big(\hat{p}\|p(\mathbf{x}|\theta)\big).$$

Indeed, we have

$$\mathsf{KL}(\hat{p}\|p(\mathbf{x}|\theta)) = \int [\log(\hat{p}(\mathbf{x})) - \log p(\mathbf{x}|\theta)]\hat{p}(\mathbf{x})\,\mathrm{d}\mathbf{x} = C + \frac{1}{n}\sum_{i=1}^{n} -\log p(\mathbf{x}_i|\theta),$$

where C is a constant that does not depend on θ.

Exercise 0.21: Is the flood gate open?

Now obviously you are thinking to replace the KL divergence with any f-divergence, hoping to obtain some generalization of MLE. Try and explain any difficulty you may run into. (We will revisit this in the GAN lecture.)

Exercise 0.22: Why KL is so special

To appreciate the uniqueness of the KL divergence, prove the following:

log is the only continuous function satisfying $f(st) = f(s) + f(t)$.

Remark 0.23: Information theory for ML

A beautiful they that connects information theory, Bayes risk, convexity and proper loss is available in (Grünwald and Dawid 2004; Reid and Williamson 2011) and the references therein.

Grünwald, P. D. and A. P. Dawid (2004). "Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory". *Annals of Statistics*, vol. 32, no. 4, pp. 1367–1433.

Reid, M. D. and R. C. Williamson (2011). "Information, Divergence and Risk for Binary Experiments". *Journal of Machine Learning Research*, vol. 12, no. 22, pp. 731–817.

Example 0.24: Linear regression as MLE

Let us now give linear regression a probabilistic interpretation, by making the following assumption:

$$Y = \mathbf{x}^\top \mathbf{w} + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Namely, the response is a linear function of the feature vector $\mathbf{x}$, corrupted by some

standard Gaussian noise, or in fancy notation: $Y \sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2)$. Given a dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1) \dots, (\mathbf{x}_n, y_n)\}$ (where we assume the feature vectors $\mathbf{x}_i$ are fixed and deterministic, unlike the responses y_i which are random), the likelihood function of the parameter $\mathbf{w}$ is:

$$L(\mathbf{w}; \mathcal{D}) = p(\mathcal{D}|\mathbf{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \mathbf{w})^2}{2\sigma^2}\right)$$

$$\hat{\mathbf{w}}_{\mathsf{MLE}} = \operatorname*{argmin}_{\mathbf{w}} \quad \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w})^2,$$

which is exactly the ordinary linear regression.

Moreover, we can now also obtain an MLE of the noise variance σ^2 by solving:

$$\begin{aligned} \hat{\sigma}^2_{\mathsf{MLE}} &= \operatorname*{argmin}_{\sigma^2} \quad \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w})^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \hat{\mathbf{w}}_{\mathsf{MLE}})^2, \end{aligned}$$

which is nothing but the average training error.

Definition 0.25: Prior

In a full Bayesian approach, we also assume the parameter θ is random and follows a prior pdf $p(\theta)$. Ideally, we choose the prior $p(\theta)$ to encode our *a priori* knowledge of the problem at hand. (Regrettably, in practice computational convenience often dominates the choice of the prior.)

Definition 0.26: Posterior

Suppose we have chosen a prior pdf $p(\theta)$ for our parameter of interest θ. After observing some data $\mathcal{D}$, our belief on the probable values of θ will have changed, so we obtain the posterior:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)\,\mathrm{d}\theta},$$

where recall that $p(\mathcal{D}|\theta)$ is exactly the likelihood of θ given the data $\mathcal{D}$. Note that computing the denominator may be difficult since it involves an integral that may not be tractable.

Example 0.27: Bayesian linear regression

Let us consider linear regression (with vector-valued response $\mathbf{y} \in \mathbb{R}^m$, matrix-valued covariate $X \in \mathbb{R}^{m \times d}$):

$$\mathbf{Y} = X\mathbf{w} + \boldsymbol{\epsilon},$$

where the noise $\boldsymbol{\epsilon} \sim \mathcal{N}_m(\boldsymbol{\mu}, S)$ and we impose a Gaussian prior on the weights $\mathbf{w} \sim \mathcal{N}_d(\boldsymbol{\mu}_0, S_0)$. As usual we assume $\boldsymbol{\epsilon}$ is independent of $\mathbf{w}$. Given a dataset $\mathcal{D} = \{(X_1, \mathbf{y}_1), \dots, (X_n, \mathbf{y}_n)\}$, we compute the posterior:

$$\begin{aligned} p(\mathbf{w}|\mathcal{D}) &\propto p(\mathbf{w})p(\mathcal{D}|\mathbf{w}) \\ &\propto \exp\left(-\frac{(\mathbf{w} - \boldsymbol{\mu}_0)^\top S_0^{-1} (\mathbf{w} - \boldsymbol{\mu}_0)}{2}\right) \cdot \prod_{i=1}^{n} \exp\left(-\frac{(\mathbf{y}_i - X_i\mathbf{w} - \boldsymbol{\mu})^\top S^{-1} (\mathbf{y}_i - X_i\mathbf{w} - \boldsymbol{\mu})}{2}\right) \\ &= \mathcal{N}(\boldsymbol{\mu}_n, S_n), \end{aligned}$$

where (by completing the square) we have

$$S_n^{-1} = S_0^{-1} + \sum_{i=1}^n X_i^\top S^{-1} X_i$$
$$\boldsymbol{\mu}_n = S_n \left(S_0^{-1}\boldsymbol{\mu}_0 + \sum_{i=1}^n X_i^\top S^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) \right).$$

The posterior covariance S_n contains both the prior covariance S_0 and the data X_i. As $n \to \infty$, data dominates the prior. Similar remark applies to the posterior mean $\boldsymbol{\mu}_n$.

We can also derive the predictive distribution on a new input X:

$$p(\mathbf{y}|X, \mathcal{D}) = \int p(\mathbf{y}|X, \mathbf{w}) p(\mathbf{w}|\mathcal{D}) \, \mathrm{d}\mathbf{w}$$
$$= \mathcal{N}(X\boldsymbol{\mu}_n + \boldsymbol{\mu}, X S_n X^\top + S)$$

The covariance $XS_nX^\top + S$ reflects our uncertainty on the prediction at X.

Theorem 0.28: Bayes classifier

Consider the classification problem with random variables $\mathbf{X} \in \mathbb{R}^d$ and $Y \in [\mathsf{c}] := \{1, \dots, \mathsf{c}\}$. The optimal (Bayes) classification rule, defined as

$$\operatorname*{argmin}_{h:\mathbb{R}^d \to [\mathsf{c}]} \Pr(Y \neq h(\mathbf{X})),$$

admits the closed-form formula:

$$h^\star(\mathbf{x}) = \operatorname*{argmax}_{k \in [\mathsf{c}]} \Pr(Y = k | \mathbf{X} = \mathbf{x}) \tag{0.1}$$
$$= \operatorname*{argmax}_{k \in [\mathsf{c}]} \underbrace{p(\mathbf{x}|Y=k)}_{\textit{likelihood}} \cdot \underbrace{\Pr(Y=k)}_{\textit{prior}},$$

where ties can be broken arbitrarily.

Proof: Let $h(\mathbf{x})$ be any classification rule. Its classification error is:

$$\Pr(h(\mathbf{X}) \neq Y) = 1 - \Pr(h(\mathbf{X}) = Y) = 1 - \mathbb{E}[\Pr(h(\mathbf{X}) = Y|\mathbf{X})].$$

Thus, conditioned on $\mathbf{X}$, to minimize the error we should maximize $\Pr(h(\mathbf{X}) = Y|\mathbf{X})$, leading to $h(\mathbf{x}) = h^\star(\mathbf{x})$.

To understand the second formula, we resort to the definition of conditional expectation:

$$\int_A \Pr(Y = k|\mathbf{X} = \mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \Pr(\mathbf{X} \in A, Y = k)$$
$$= \Pr(\mathbf{X} \in A | Y = k) \Pr(Y = k)$$
$$= \int_A p(\mathbf{x}|Y=k) \Pr(Y=k) \, \mathrm{d}\mathbf{x}.$$

Since the set A is arbitrary, we must have

$$\Pr(Y = k|\mathbf{X} = \mathbf{x}) = \frac{p(\mathbf{x}|Y=k)\Pr(Y=k)}{p(\mathbf{x})}.$$

(We assume the marginal density $p(\mathbf{x})$ and class-specific densities $p(\mathbf{x}|Y=k)$ exist.) ■

In practice, we do not know the distribution of $(\mathbf{X}, Y)$, hence we cannot compute the optimal Bayes classification rule. One natural idea is to estimate the pdf of $(\mathbf{X}, Y)$ and then plug into (0.1). This approach however does not scale to high dimensions and we will see direct methods that avoid estimating the pdf.

It is clear that the Bayes error (achieved by the Bayes classification rule) is:

$$\mathbb{E}\big[1 - \max_{k\in[\mathsf{c}]} \Pr(Y = k|\mathbf{X})\big].$$

In particular, for $\mathsf{c} = 2$, we have

$$\text{Bayes error} = \mathbb{E}\big[\min\{\Pr(Y = 1|\mathbf{X}), \Pr(Y = -1|\mathbf{X})\}\big].$$

Remark 0.29: Generalized Bayes Theorem (e.g., Shiryaev 2016, p. 272)

Let $\mathcal{G}$ be a sub-σ-field. By definition, the conditional expectation is simply the Radon-Nikodym derivative:

$$\mathbb{E}[f(Y)|\mathcal{G}] = \frac{\mathrm{d}Q}{\mathrm{d}P_{\restriction\mathcal{G}}}, \quad \text{where} \quad Q(A) := \int_A f(Y)\,\mathrm{d}P,\ \forall A \in \mathcal{G}.$$

Suppose there exists a jointly measurable $p(\omega|y)$ and (σ-finite) measure $\lambda(\omega)$ so that

$$\boxed{\forall A \in \mathcal{G},\ P(A|Y = y) = \int_A p(\omega|y)\,\mathrm{d}\lambda(\omega).}$$

(It follows that $P(A|Y)$ admits a regular version, i.e., RHS.) Thus,

$$P(A) = \mathbb{E}[P(A|Y)] = \int \left[\int_A p(\omega|y)\,\mathrm{d}\lambda(\omega)\right] \mathrm{d}P_Y(y)$$

$$Q(A) = \mathbb{E}[f(Y)[\![A]\!]] = \mathbb{E}\Big[f(Y)\cdot\mathbb{E}[[\![A]\!]|Y]\Big] = \int f(y)P(A|Y = y)\,\mathrm{d}P_Y(y) = \int f(y)\left[\int_A p(\omega|y)\,\mathrm{d}\lambda(\omega)\right]\mathrm{d}P_Y(y).$$

Therefore,

$$\boxed{\mathbb{E}[f(Y)|\mathcal{G}](\omega) = \frac{\mathrm{d}Q/\,\mathrm{d}\lambda}{\mathrm{d}P_{\restriction\mathcal{G}}/\,\mathrm{d}\lambda} = \frac{\int f(y)\cdot p(\omega|y)\,\mathrm{d}P_Y(y)}{\int p(\omega|y)\,\mathrm{d}P_Y(y)}.}$$

Letting $f(Y) = [\![Y \in B]\!]$ we obtain the Bayes formula for the posterior distribution, which automatically admits a regular version (i.e., RHS)!

We also record the following change-of-measure formula from Shiryaev (2016, Theorem 6, p. 275):

$$\text{If } Q \ll P, \quad \boxed{\mathbb{E}_Q[X|\mathcal{G}] = \frac{\mathbb{E}_P[X\frac{\mathrm{d}Q}{\mathrm{d}P}|\mathcal{G}]}{\mathbb{E}_P[\frac{\mathrm{d}Q}{\mathrm{d}P}|\mathcal{G}]}} \quad (Q - a.s.). \tag{0.2}$$

Shiryaev, A. N. (2016). "Probability-1". Springer.

Definition 0.30: Sufficiency

We call a sub-σ-field $\mathcal{G}$ sufficient for the family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if there exist (versions of) conditional probabilities $P_\theta(\cdot|\mathcal{G})(\omega)$ that no longer depend on θ, i.e., for all $\theta \in \Theta$ and $A \in \mathcal{F}$,

$$P_\theta(A|\mathcal{G})(\omega) = P(A|\mathcal{G})(\omega), \quad (P_\theta - a.s.).$$

For a dominated family, $\mathcal{G}$ is sufficient iff the Radon-Nikodym derivatives factorize:

$$p_\theta(\omega) := \frac{\mathrm{d}P_\theta}{\mathrm{d}\lambda} = g_\theta(\omega) \cdot h(\omega), \quad (\lambda - a.s.)$$

where g_θ is $\mathcal{G}$-measurable. Indeed, applying (0.2) we obtain

$$P_\theta(A|\mathcal{G}) = \frac{\mathbb{E}_\lambda[[\![A]\!] g_\theta h|\mathcal{G}]}{\mathbb{E}_\lambda[g_\theta h|\mathcal{G}]} = \frac{\mathbb{E}_\lambda[[\![A]\!] h|\mathcal{G}]}{\mathbb{E}_\lambda[h|\mathcal{G}]},$$

which does not depend on θ. See Billingsley (1995, Theorem 34.6) for a full proof of the converse and Borovkov (1998, Chapter 2) for the construction of minimal sufficient σ-field.

Billingsley, P. (1995). "Probability and Measure". 3rd. Wiley.
Borovkov, A. A. (1998). "Mathematical Statistics". CRC Press.

Exercise 0.31: Cost-sensitive classification (Elkan 2001)

Cost-sensitive classification refers to the setting where making certain mistakes is more expensive than making some other ones. Formally, we suffer cost c_{ij} when we predict class i while the true class is j. We may of course assume $c_{ii} \equiv 0$. Derive the optimal Bayes rule.

Elkan, C. (2001). "The Foundations of Cost-Sensitive Learning". In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 973–978.

Exercise 0.32: Bayes estimator

Let $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function that compares our prediction $\hat{\mathbf{y}}$ with the groundtruth $\mathbf{y}$. We define the Bayes estimator as:

$$\min_{f:\mathcal{X}\to\hat{\mathcal{Y}}} \mathbb{E}\ell(f(\mathbf{X}), \mathbf{Y}).$$

Can you derive the formula for the Bayes estimator (using conditional expectation)?

Definition 0.33: Maximum a posteriori (MAP)

Another popular parameter estimation algorithm is the MAP that simply maximizes the posterior:

$$\begin{aligned}\theta_{\mathsf{MAP}} &:= \operatorname*{argmax}_{\theta\in\Theta} \; p(\theta|\mathcal{D}) \\ &= \operatorname*{argmin}_{\theta\in\Theta} \quad \underbrace{-\log p(\mathcal{D}|\theta)}_{\text{negative log-likelihood}} \quad + \quad \underbrace{-\log p(\theta)}_{\text{prior as regularization}}\end{aligned}$$

A strong (i.e. sharply concentrated, i.e. small variance) prior helps reducing the variance of our estimator, with potential damage to increasing our bias (see Definition 0.10) if our *a priori* belief is mis-specified, such as stereotypes ☺.

MAP is *not* a Bayes estimator, since we cannot find an underlying loss ℓ for it.

Example 0.34: Ridge regression as MAP

Continuing Example 0.24 let us now choose a standard Gaussian prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\lambda}\mathbb{I})$. Then,

$$\hat{\mathbf{w}}_{\mathsf{MAP}} = \operatorname*{argmin}_{\mathbf{w}} \quad \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \mathbf{w})^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 - \frac{d}{2}\log\lambda,$$

which is exactly equivalent to ridge regression. Note that the larger the regularization constant λ is, the smaller the variance of the prior is. In other words, larger regularization means more determined prior information.

Needless to say, if we choose a different prior on the weights, MAP would yield a different regularized linear regression formulation. For instance, with the Laplacian prior (which is more peaked than the Gaussian around the mode), we obtain the celebrated Lasso (Tibshirani 1996):

$$\min_{\mathbf{w}} \quad \frac{1}{2\sigma^2}\|X\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1.$$

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288.

Theorem 0.35: Bayes rule arose from optimization (e.g. Zellner 1988)

Let $p(\theta)$ be a prior pdf of our parameter θ, $p(\mathcal{D}|\theta)$ the pdf of data $\mathcal{D}$ given θ, and $p(\mathcal{D}) = \int p(\theta)p(\mathcal{D}|\theta)\,\mathrm{d}\theta$ the data pdf. Then,

$$p(\theta|\mathcal{D}) = \operatorname*{argmin}_{q(\theta)} \quad \mathsf{KL}\big(p(\mathcal{D})q(\theta) \parallel p(\theta)p(\mathcal{D}|\theta)\big), \tag{0.3}$$

where the minimization is over all pdf $q(\theta)$.

Proof: KL is nonnegative while the posterior $p(\theta|\mathcal{D})$ already achieves 0. In fact, only the posterior can achieve 0, see Exercise 0.14. ■

This result may seem trivial at first sight. However, it motives a number of important extensions:

- If we restrict the minimization to a subclass $\mathcal{P}$ of pdfs, then we obtain some KL projection of the posterior $p(\theta|\mathcal{D})$ to the class $\mathcal{P}$. This is essentially the so-called variational inference.
- If we replace the KL divergence with any other f-divergence, the same result still holds. This opens a whole range of possibilities when we can only optimize over a subclass $\mathcal{P}$ of pdfs.
- The celebrated expectation-maximization (EM) algorithm also follows from (0.3)!

We will revisit each of the above extensions later in the course.

Zellner, A. (1988). "Optimal Information Processing and Bayes's Theorem". *The American Statistician*, vol. 42, no. 4, pp. 278–280.